Identification and characterization of rearrangements in the vervet monkey genome

by

AmanPreet Badhwar

Department of Human Genetics
McGill University, Montreal

August 2006

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of

Master of Science

Your file Votre référence
ISBN: 978-0-494-32815-6
Our file Notre référence
ISBN: 978-0-494-32815-6

Abstract

Several mechanisms can lead to the reorganization of genomes during speciation, including centromere repositioning, new centromere emergence or other chromosomal rearrangements. Using a comparative karyotype approach, I determined that the vervet genome contains at least 12 evolutionary young centromere locations.

To study the evolutionary dynamics of centromere formation, I identified and validated the alpha-satellite repeat as a centromere-specific marker in the vervet using comparative genomics, sequence analysis and hybridization screening. I developed criteria to infer the position of vervet bacterial artificial chromosome (BAC) inserts based on alpha-satellite monomer content. In a complementary approach, I demarcated the pericentromeric boundaries in human and identified vervet BAC clones that mapped orthologously to these regions.

In addition to centromeric analyses, I developed methodologies to detect other genome rearrangements, in particular vervet deletion/human insertion and vervet translocation events. The tools and approaches developed in this thesis will prove useful in cataloguing additional vervet genome rearrangements.

Résumé

Plusieurs mécanismes sont susceptibles d'engendrer une réorganisation du génome lors d'une spéciation, comme par exemple un repositionnement de centromère, l'apparition de nouveaux centromères et autres réarrangements chromosomiques. En utilisant une méthode basée sur la comparaison des caryotypes, j'ai déterminé que le génome du Vervet (singe vert) contient un minimum de 12 positions centromériques récemment apparues au cours de l'évolution.

Pour ce faire, j'ai utilisé une approche combinant des méthodes de génomique comparative, d'analyse de séquences et de criblage par hybridation, ce qui m'a permis d'identifier et de valider un motif répété alpha-satellite pouvant servir de marqueur spécifique pour les centromères. De plus, j'ai développé des critères pour déterminer la position des BAC (chromosomes artificiels de bactérie) du Vervet en fonction de leur composition en monomères alpha-satellites. En complément, j'ai défini les limites des péri-centromères humains et identifié les BAC du Vervet correspondant à ces régions.

Par ailleurs, j'ai également développé des méthodes afin de détecter d'autres types de réarrangements chromosomiques, notamment les translocations chez le Vervet ainsi que les portions du génome délétées chez le Vervet ou insérées chez l'humain. Ces nouveaux outils faciliteront le catalogage des autres réarrangements génomiques chez le Vervet.

Table of Contents

Table of Figures

Table of Abbreviations

ALR: alpha-satellite repeat

BAC: Bacterial Artificial Chromosome

BACPAC: Bacterial Artificial Chromosome Plasmid Artificial Chromosome

CHORI: Children's Hospital Oakland Research Institute

DNA: DeoxyriboNucleic Acid

E. coli: Escherichia coli

FISH: Fluorescent In Situ Hybridization

Gb: Giga bases

kb: Kilo bases

Mb: Mega bases

mya: million years ago

NCBI: National Center for Biotechnology Information

PCR: Polymerase Chain Reaction

RNA: RiboNucleic Acid

SNP: Single Nucleotide Polymorphism

UCLA/VA: University of California, Los Angeles/Veterans Administration

UCSC: University of California Santa Cruz

YAC: Yeast Artificial Chromosome

Acknowledgements

The work presented in this thesis would have been impossible without the contributions of many people.

First and foremost I would like to thank Dr. Ken Dewar, my supervisor, for proposing a project that introduced me to the field of comparative genomics and bioinformatics and for providing an environment conducive to research and learning. Dr. Dewar has been a source of constant support and inspiration for each phase of this challenging yet rewarding journey. Under his guidance I have been able to hone my ability to pose scientifically interesting questions and designing successful experiments that provide the answers. Overall, I am very grateful to him for allowing me the freedom to try out different avenues of analyses yet keeping me grounded at all stages of the project.

I would also like to thank the members of my advisory committee, Drs. Roberta Palmour and Mathieu Blanchette. I am indebted to them for their advice and thoughts with regards to my research work. I am grateful to Dr. Blanchette for asking insightful bioinformatics questions that helped strengthen various aspects of my research work. I am also thankful to Dr. Palmour for shedding light on the various aspects of vervet monkey genomics and for her words of encouragement with regards to my project.

From the Department of Human Genetics I would like to thank Laura Benner (our former graduate program coordinator) and Fran Langton for their help in sorting out my complicated application and for putting me in touch with Dr. Dewar's lab. I would also

like to thank Kandace Springer, our present graduate program coordinator, for her wonderful ability to solve all my administrative dilemmas and for being supportive in times of need.

I would also like to thank our entire bioinformatics group (Vince Forgetta, Kevin McKee, Jessica Wasserscheid, Emmanuel Mongin and Raman Minhas) for introducing me to this field. In particular, I am grateful to Vince and Kevin for providing me with the various "Introduction to....." books, and for teaching me how to use the different bioinformatics softwares. I am also very thankful to Kevin for answering all my databasing questions and to Jessica for last-minute help in getting things done under the wire.

I would also like to take this opportunity to thank Gary Leveque, Xiaolan Zhang and Carol Dore for teaching me all about the BAC end sequencing platform. A special thanks to Gary for agreeing to help me with the onerous job of tracking and timing every step the robot executed. I definitely needed an extra pair of hands with the timer starts and stops.

This journey would have been a lot harder without the friendship, support, and guidance of the administrative staff, professors, postdoctoral fellows and graduate students at the McGill Genome Centre. I would thus like to thank all of them for standing by me through all the good and bad times. In particular, I would like to thank Dr. David Serre, Alexandra Chloe Villani and Guillaume Paré for translating my thesis abstract to French. I would also like to thank Dr. David Serre for his insightful comments during my many brainstorming sessions and for translating my thesis abstract to French. He also

made sure that our entire team had a social life outside the lab, in particular the many trips to Thompson house as well as several late night board game nights. Thanks to these little interruptions I always started with a fresh mind the next day.

Last but not least I would like to thank my better half, Sridar Narayanan for his love and encouragement throughout this endeavour and for especially helping me get through the final thesis-writing phase. I would also like to thank my parents and brother who despite being all the way in India have been a constant source of support and inspiration.

Special Recognitions

In this section, I would like to thank certain colleagues for helping with specific aspects of my research.

Chapter Three

I would like to thank Vinceno Forgetta for showing me how to efficiently use the various features of the UCSC Browser, how to generate annotated chromosomes and how to fine-tune RepeatMasker outputs.

In addition, many thanks to Xiaolan Zhang for guiding me through the probe design process, as well as for helping me conduct the hybridization screen experiment. I learnt a lot from her expertise.

I would also like to recognize Kevin McKee for scripting the Filter Interpreter application. This tool allowed me to identify the vervet BACs with alpha satellite content

present in the high-density filter (Filter 1) that was used for the hybridization screen experiment.

Chapter Four

Many thanks to Kevin McKee for providing database support, and for constantly fine tuning his scripts to satisfy my curiosities. Without his help it would have been a long and tedious process to combine the sequence analyses and mapping data.

Chapter Five

I would like to thank Gary Leveque for providing me with the protocol for gel electrophoresis and for always being on hand if I had any questions while conducting the experiments.

In addition, a special thanks to Jessica Wasserscheid for customizing the Vervet Monkey Genomics website so that the layout of the orthologous mapping (vervet with respect to human) results would be immediately useful for my analyses. This saved me precious time while using BAC end mapping data for identifying regions of genome rearrangements.

I would like to thank Kevin McKee for collaborating with me in the design of the visualization tool used to detect clustering of discordant inter-chromosomal vervet BACs. Without his diligent after-hours scripting the tool would have remained in the realm of fiction.

Contributions to original knowledge

My master's thesis work has resulted in several original contributions to scientific knowledge, in particular to the fields of primate genomics and evolution. These contributions comprise of:

- The identification of 12 evolutionary young centromeric locations in the vervet genome.
- The development of two methodologies to identify vervet BAC clones belonging to centromeric regions.
- The design of two overgo probe sets that can be used to identify vervet bacterial artificial clones (BAC) with alpha-satellite content.
- The discovery that in addition to ancestral alpha-satellite (ALR) monomers the vervet genome also consists of evolutionary young sub-families of ALR monomers.
- Preliminary evidence that the ALR monomer content of the functional domains of vervet centromeres differ from that of the human.
- The design of a tool that provides rapid identification of vervet inter-chromosomal translocation events.

Chapter One

Introduction

Model organisms in biomedical research are defined as "surrogates for a human being or a human biological system, and are used to understand normal and abnormal function from gene to phenotype to provide a basis for preventive or therapeutic intervention in human diseases" (Rogatcheva et al. 2004). A multitude of viruses, prokaryotes and eukaryotes are used as model organisms in biomedical research each year. The conservation of the biological mechanism being studied is usually the determining factor for pairing a model organism with a specific study. Other factors that influence the model organism selection process include (Bier and McGinnis 2003):

- ease of procuring the organism
- size of the organism
- ease of experimental manipulations
- availability of genomic data
- evolutionary relationship to man
- cost effectiveness

1.1 Nonhuman primates in research

Of the numerous model organisms used in research, nonhuman primates are considered unique and "constitute irreplaceable high fidelity models" for studying a

variety of human health issues, ranging from diseases and disorders, to potential therapies and preventive strategies (Carlsson et al. 2004). Nonhuman primates are phylogenetically the nearest kin to man and share a high degree of similarity with humans in terms of morphology, physiology, behaviour, and cognitive capabilities (VandeBerg and Williams-Blangero 1996; Palmour et al. 1997; Eichler and DeJong 2002; Chan 2004). Due to this shared ancestry nonhuman primates make invaluable biomedical models, in particular for the study of complex human traits (for example cognitive ability and personality) that tend to lack counterparts in phylogenetically distant animal models (Palmour et al. 1997; Bellino and Wise 2003; Deaner et al. 2005; Nadon 2006).

The total number of nonhuman primates used in research globally is estimated to be in the vicinity of 100,000-200,000 animals annually and accounts for less than one percent of all animals used (Hau et al. 2000; Carlsson et al. 2004). Traditionally, nonhuman primates have been used most commonly in microbiology, neuroscience and biochemistry research (Carlsson et al. 2004; Hau and Schapiro 2006). Recent years have also seen an increase in the usage of nonhuman primate models for genetic studies. This increase has primarily resulted due to the availability of primate reference genome maps; human, chimpanzee and rhesus monkey (Lander et al. 2001; Mikkelsen et al. 2005)(www.hgsc.bcm.tmc.edu/projects/rmacaque). The observation of conserved synteny between the primate genome maps generated to date has provided improved opportunities for genomic data from humans to be used in the analysis of nonhuman primate species and vice versa (2005; Milosavljevic et al. 2005; Murphy et al. 2005; Rogers et al. 2006). Table 1 provides examples of some of the major biomedical advances to date that have been aided by research on nonhuman primates.

Timeline	Biomedical areas of development		
1900s	Blood & plasma components	pellagra	
1920s		typhoid fever	
1930s	Mumps virus Anesthesia agents Neuromuscular blocking agents		
1940s	Rh factor	safer blood transfusions rheumatoid arthritis	
1950s	Chlorpromazine & other tranquilizers	cancer chemotherapy	Polio Yellow fever
1960s	Mapping of heart's connections to arteries Therapeutic use of cortisone	corneal transplants	German measles
1970s	Tumor viruses interact with genetic material Slow nervous system viruses	leprosy Restoration of brain blood supply	
1980s	Visual processing in the brain Psychophysiological co-factors in anxiety & phobias Cyclosporine & anti-rejection drugs Primate model for Parkinson's disease Rhesus monkey model for AIDS	Chemotherapy related malnutrition Congenital cataracts & "lazy eye" in children Heart & lung transplants for cardiopulmonary hypertension Taurine added to infant formulas for normal retinal development	Hepatitis B
1990s	Mood regulation by estrogen Lead toxicity One-dose transplant drug prevents organ rejection Dangers of moderate alcohol levels in pregnancy Mechanisms & disorders of puberty Primate embryonic stem cells shed light on reproduction & genetic disorders Ecstasy causes long-term brain damage Primate model for diabetes Regenerative mechanisms in mature brain Characterizing emerging infectious diseases Use of rhesus & cynomolgus monkey kidneys for diagnosing influenza	Post-menstrual, -partum and -menopausal depression Control of intimal hyperplasia Parent to child lung transplants for cystic fibrosis	Anthrax
2000s	Gene therapy for monkeys with Parkinson Primate model to study effects of malaria in pregnant women & offspring Characterization of cyclospora	Anti-viral drugs for HIV treatment Health benefits of dietary changes Human embryonic stem cell therapy	

Table 1: Major biomedical advances aided via nonhuman primate research

This table was adapted from the website http://pin.primate.wisc.edu/research/pibr/contribs.html

Among the nonhuman primates, the apes (Family *Hominidae*) are the closest relatives to humans (Figure 1), followed by the Old World monkeys, the New World monkeys, the prosimians, and the tarsiers (Goodman 1999). The apes and Old World monkeys both belong to the infraorder *Catarrhini* and are distinguished from the New World monkeys (infraorder *Platyrrhini*) by their "characteristically larger brains, greater manual dexterity, and improved hand-eye coordination" (Schwimmer 1998).

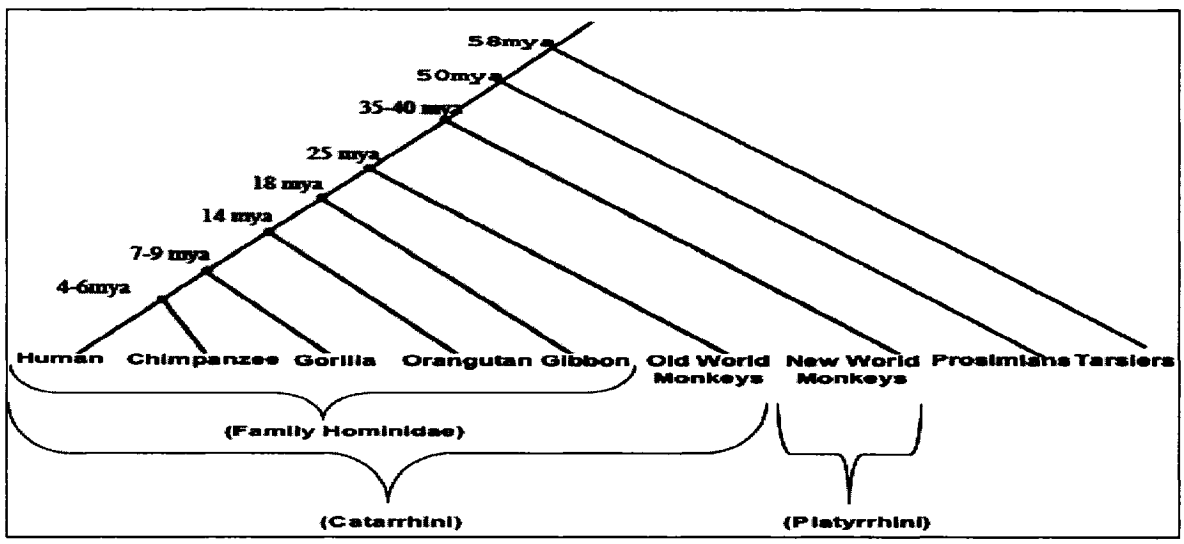

Figure 1: Primate phylogenetic tree showing divergence times

Abbreviations: mya; million years ago. The diagram has been adapted from (Goodman 1999) and (www.genome.gov/Pages/Research/Sequencing/SeqProposals/GibbonFINALseq.pdf).

The Old World monkeys constitute the predominant nonhuman primate species used in laboratory based research, followed by the New World monkeys and the apes (Carlsson et al. 2004). An admixture of both arboreal and terrestrial species, the Old World monkeys occupy a wide variety of ecosystems in their native habitats of Africa and Asia (Cawthon Lang 2006). There are at least 135 Old World monkey species identified to date belonging to the two subfamilies of *Cercopithecinae* and *Colobinae*. Of these, approximately 5-10 species consisting primarily of the macaques (*Macaca*), the baboons

(*Papio*) and the vervets (*Chlorocebus*) are used extensively for biomedical studies (Carlsson et al. 2004). Table 2 provides a list of these frequently used Old World monkey species.

[redacted]	[redacted]
macaque (*Macaca*)	rhesus macaque (*M. mulatta*), cynomolgus macaque (*M. fascicularis*), pigtailed macaque (*M. nemestrina*)
baboon (*Papio*)	yellow baboon (*P. cynocephalus*), olive baboon (*P. anubis*), guinea baboon (*P. paplo*), chacma baboon (*P. ursinus*), hamadryas baboon (*P. hamadryas*)
patas monkey (*Erythrocebus*)	*Erythrocebus patas*
vervet monkey (*Chlorocebus*)	savanna monkey (*Ch. aethiops*), malbrouck (*Ch. cynosuros*), djam-djam (*Ch. djamdjamensis*), vervet (*Ch. pygerythrus*), green monkey (*Ch. sabaeus*) and tantalus monkey (*Ch. tantalus*)

Table 2: Old World monkey species most used for nonhuman primate research

Abbreviations: *M.*; Macaca, *P.*; Papio, *Ch.*; Chlorocebus

Note: Vervet monkeys were recently moved from the genus *Cercopithecus* to a new genus *Chlorocebus*. To date, six different species of vervet monkeys have been identified. However, in the literature, regardless of the specific species used, vervets are commonly referred to as *Chlorocebus aethiops* (*Ch. aethiops*) or by their former genus, *Cercopithecus aethiops* (*C. aethiops*). Due to this practice, commenting on the usage of any specific vervet species is difficult (Carlsson et al. 2004; Cawthon Lang 2006).

A recent survey of nonhuman primate research (live animals and biological materials including commercial tissue culture cells) identified the vervets as the primate species most used in biomedical research (Carlsson et al. 2004).

1.2 Vervet monkey as a biomedical model

Six different subspecies of vervet monkeys have been identified to date (Cawthon Lang 2006): *Chlorocebus aethiops aethiop* (grivet or savanna monkey), *Chlorocebus aethiops cynosuros* (malbrouck), *Chlorocebus aethiops djamdjamensis* (Bale Mountains vervet or djam-djam), *Chlorocebus aethiops pygerythrus* (vervet monkey), *Chlorocebus*

aethiops sabaeus (green monkey) and *Chlorocebus aethiops tantalus* (tantalus monkey). Omnivorous and tolerant of climate and ecologies, the vervets are native to Africa and inhabit large parts of sub-Saharan Africa (Palmour et al. 1997). In addition, *Chlorocebus aethiops sabaeus* is also found on the Cape Verde islands and on several Caribbean islands in the West Indies (Cawthon Lang 2006). The green monkey was introduced (as pets or as items to be traded) to the Caribbean islands of St. Kitts, Nevis, and Barbados in the late 1600s by ships involved in the slave trade (Van der Kuyl et al. 1996).

1.2.1 Attractive vervet attributes for biomedical research

With respect to biomedical research, the vervet has several features that make it an attractive nonhuman primate model (Palmour et al. 1997; Ervin and Palmour 2003).

- **Nonendangered:** The vervet is highly adaptable and flourishing in most habitats, so much so that it is referred to as 'the weed monkey of Africa' (Fedigan and Fedigan 1988). In the Caribbean, where it lacks natural predators, the vervet population has grown unchecked and is considered a serious agricultural threat (Boulton et al. 1996).
- **Small in size:** Slightly smaller than the macaque and considerably smaller than the baboon, the vervets are more economical in regards to housing space requirements (Fairbanks 2003).
- **Temperamentally malleable:** The vervet is more tractable temperamentally compared to the macaque and the baboon making it easier to handle.
- **Highly fertile:** Vervets are also relatively easy to breed in captivity. The female vervet has a higher fecundity rate than the macaque producing up to three infants in two years. (Fairbanks 2003).

- **Free of herpes B virus:** Unlike the macaques, vervets from both Africa and the Caribbean are free of herpes B virus (Palmour et al. 1997). Moreover, Caribbean vervets are also free of other significant pathogenic viruses including the African green monkey variant of simian immunodeficiency virus, Cercopithicine herpesvirus 2 (SA-8), and Marburg virus that infested the African vervet and other Old World monkey populations (Palmour et al. 1997; Fairbanks 2003).
- **Models human disorders:** The vervet is used as a biomedical model for several human disorders (Garcia-Castells et al. 1989; Ervin et al. 1990; Martin et al. 1990; Liberini et al. 1993; Ervin and Palmour 2003; Lemere et al. 2004).

Spontaneous	Induced
• Hypertension	• Parkinson's disease
• Metabolic syndrome (X)	• Multi-infarct dementia
• Polycystic ovarian disease	• Allotransplant GvH syndrome
• Alcohol abuse & fetal alcohol syndrome	• Xenotransplant GvH
• Anxiety disorders	• Estogen-induced uterine CA
• Mother-fetus dyads (e.g., gor vaccine teratology)	• Cutaneous leishmaniasis
• Cataracts	
• Alzheimer's disease	
• Menstrual disorders	

Note: This table has been adapted from (Ervin and Palmour 2003)

- **Easily obtained:** Several large research facilities house and export vervets. These include the St. Kitts Biomedical Research Foundation, Caribbean Primate Research Laboratories, Barbados Primate Research Center, New Iberia Research Center and University of California, Los Angeles/Veterans Administration (UCLA/VA) Vervet Research Colony. Smaller vervet colonies can be found at Wake Forest University and the University of Texas at Austin (Fairbanks 2003).
- **Resonably priced:** The vervet is competitively priced in comparison to other nonhuman primate species (Palmour et al. 1997; Fairbanks 2003).

1.2.2 Availability of vervet genetic resources

Many studies using the vervet entail the determination of correlations between phenotypes and genetic loci. Such studies are primarily possible due to the existence of various genetic resources that are available for use with the vervet. Pedigreed multi-generation vervet monkey colonies maintained by McGill University/St. Kitts and the UCLA/VA research groups, are two examples such resource (Palmour et al. 1997; Bailey et al. 2001a; Ervin and Palmour 2003). Vervet pedigreed colonies serve to (i) establish the mode of inheritance (eg. dominant, recessive or other) of various traits; (ii) permit the controlled propagation of phenotypes under investigations; (iii) permit genotype-phenotype correlations of simple and complex traits; (iv) allow for the successful completion of longitudinal studies; and (v) provide a genetic mapping resource.

A second resource under development is a combined microsatellite and single nucleotide polymorphism (SNP) genetic map compatible with both the McGill and UCLA pedigreed vervet colonies (NIH 1R01RR016300-01A1). Availability of such a map will support both genome-wide and fine-scale genetic mapping studies in the vervet. Our laboratory is further contributing to these genetic resources by generating a BAC-based physical map of the vervet (*Chlorocebus aethiops sabaeus*). Development and use of a vervet physical map will help researchers in establishing homology between genes present in the vervet, with those present in human, other nonhuman primate species, and more distant lineages such as the mouse and pufferfish. This will lead to improved opportunities for the identification of new genes, "either disease specific or functional genes that influence normal anatomical, physiological and behavioural variation, as well

as genes that are targets for therapeutic intervention" (Palmour et al. 1997; Rogers and VandeBerg 1998).

1.3 Vervet monkey as a model of genome evolution

Comparison of multiple ape and Old World monkey physical maps can lead to the identification of diverged genomic sequences that may be responsible for species-specific traits (Frazer et al. 2003; Enard and Paabo 2004). The human, the chimpanzee, and the rhesus genomes are the only primate genomes sequenced and assembled to date (*http://genome.ucsc.edu/*). A vervet physical map will complement that of the rhesus in identifying evolutionary features distinct between Old World monkeys and apes. At present, the rhesus genomic sequence is used as an outgroup for distinguishing sequences and genes specific to the great ape lineage from those specific to either human or chimp. However, these analyses do not allow us to determine whether a particular feature in the rhesus genome is common to all Old World monkeys, or is limited to a particular species or individual. Having the vervet physical map will enable the identification of sequences and/or reorganization truly unique to humans from evolutionary recent features shared with other living primates (Dubchak and Frazer 2003; Frazer et al. 2003; Nobrega and Pennacchio 2003).

Another important aspect of the vervet physical map is that it provides a unique opportunity for the study of genome evolution, particularly for chromosome fission. Compared to the apes and other Old World monkey genomes, the vervet genome (in

particular, *Chlorocebus aethiops sabaeus*) has undergone major remodelling (Figure 2), resulting in a high karyotype number of 2n=60 (Finelli et al. 1999; Goodman 1999).

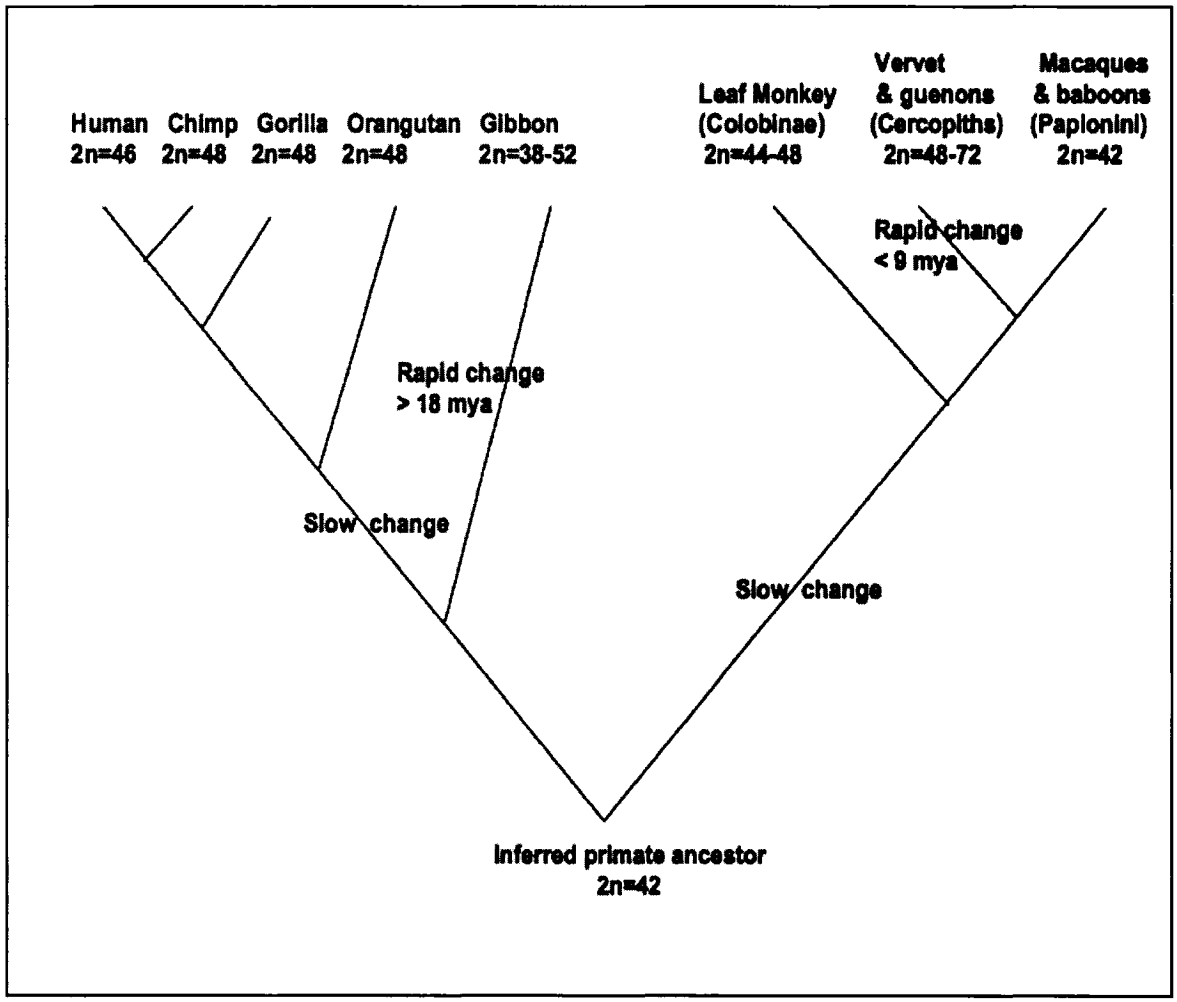

Figure 2: A partial primate phylogenetic tree

This phylogenetic tree has been adapted from (Jones et al. 1992). The main goal of this tree is to show that most apes and old world monkeys have similar chromosome numbers (2n = 42-48). The increase in chromosome numbers observed in some species of gibbons (2n = 38-52) and Cercopiths (2n = 48-72) is believed to be a consequence of extensive genome remodeling. Based on the estimated divergence times, reorganization in Cercopith genomes is considered more recent (< 9 mya) to that observed in the gibbon lineage (> 18 mya). Abbreviations: mya; million years ago.

In contrast, the karyotypes of man and the Pongidae (chimpanzee, bonobos, gorilla, and orangutan) "differ by a small number of rearrangements, mainly pericentric inversions and one fusion that reduced the chromosome number from 48 in the Pongidae to 46 in

man" (De Grouchy 1987). Similarly, the Papionini and Colobinae karyotype numbers range from 2n = 42-48 (Stanyon et al. 1988; Jauch et al. 1992).

Overall in the primate lineage, apart from the vervet, the gibbon is the only other species with an extensively rearranged karyotype (Muller et al. 1997; Muller and Wienberg 2001). However, compared to the gibbon, the reorganization of vervet genome is more recent (Figure 2), as the Cercopithecina/Papionina split is estimated at around 9 mya and the gibbon split occurred approximately 18 mya (Goodman 1999). The increased number of chromosomes in the vervet lineage hints that its genome carries a blend of ancestral and new centromeres. A physical map of the vervet will allow for the confirmation and delineation of the exact chromosomal breakpoint positions, as well as shed light on the evolution of new centromeres, pericentromeric regions, and telomeres.

1.4 Vervet physical map project

There are two general strategies for sequencing large genomes; whole genome shotgun sequencing and clone-based shotgun sequencing. Shotgun sequencing (Sanger et al. 1980; Sanger et al. 1982; Batzoglou et al. 1999) is achieved by 'breaking a target into random fragments, sequencing these fragments, and reconstructing the full sequence from these pieces' (Sanger et al. 1980; Sanger et al. 1982; Batzoglou et al. 1999). In the whole genome method, the entire genome is targeted for fragmentation all at once. Contrastingly, the clone-based method requires an initial collection of large overlapping clones with known locations covering the genome. Since the clone-based method fragments and assembles each individual clone, large-scale mis-assemblies, which can occur in the whole-genome method, are averted (Batzoglou et al. 1999).

1.4.1 BAC Cloning System

Due to the size of mammalian genomes (3.0-3.2 billion base pairs), the availability of large (150-1000 kb) insert DNA libraries is crucial for clone-based sequencing. Such libraries are usually generated by yeast artificial chromosome (YAC) or bacterial artificial chromosome (BAC) cloning systems. The vector used to clone DNA fragments in the BAC system is derived from an endogenous F-factor plasmid found in the bacterium Escherichia coli (*E. coli*) which contains genes for strict copy number control and unidirectional origin of DNA replication (Shizuya et al. 1992). BAC libraries can be generated by ligating size-selected restriction digested DNA with the BAC vector followed by electroporation into *E. coli.*

In recent years, BAC libraries have gained widespread use, with 164 different vertebrate libraries now available from BACPAC Resources (http://bacpac.chori.org/) alone. This is mainly due to the ease of library generation and insert manipulation associated with the system. Unlike YACs, the BAC system is usually free of problems such as low yields of insert DNA and clone rearrangements, deletions, and chimerisms (Woo et al. 1994; Venter et al. 1996). The latter is primarily because BACs are maintained as single copy plasmids in *E. coli* and exclude other BAC plasmids from replicating in the same host cell (Kim et al. 1992; Shizuya et al. 1992; Woo et al. 1994). BAC libraries can be screened by polymerase chain reaction (PCR) methods or by hybridization approaches, enabling researchers to obtain clones of their regions of interest even in the absence of large-scale genome sequencing initiatives.

1.4.2 Vervet BAC End Sequencing

BAC end sequencing involves the sequencing of the two ends of a BAC insert using primers located in the two vector arms (Venter et al. 1996). The end result is the generation of two short stretches of DNA ranging from 500 – 800 bp in length from the two extremities of a BAC insert. BAC end sequencing has become a routine activity in the complete sequencing of large genomes, either for establishing clone tiling paths (in clone based projects) or for providing long range physical links (in whole genome shotgun assemblies).

At present, our laboratory is in the process of generating a physical map framework for the vervet genome. This project requires (i) large-scale end sequencing of >200,000 clones of the CHORI-252 male vervet BAC library generated at BACPAC Resources (Children's Hospital Oakland Research Institute) and (ii) the alignment of vervet BAC clone ends (using bioinformatics approaches) to syntenic regions of the human, chimpanzee, and rhesus genome assemblies. The project is estimated to take three years, and the complete map is expected to provide approximately 8-10 fold coverage of the vervet genome.

The vervet map resource will positionally link genetic markers to neighbouring genes, and will provide information on the availability of BAC clones and BAC clone paths spanning genomic regions of interest. Having started in October 2005, the project is still in its early stages. The vervet BAC end sequencing conducted to date provides approximately 0.75X coverage of the vervet monkey genome.

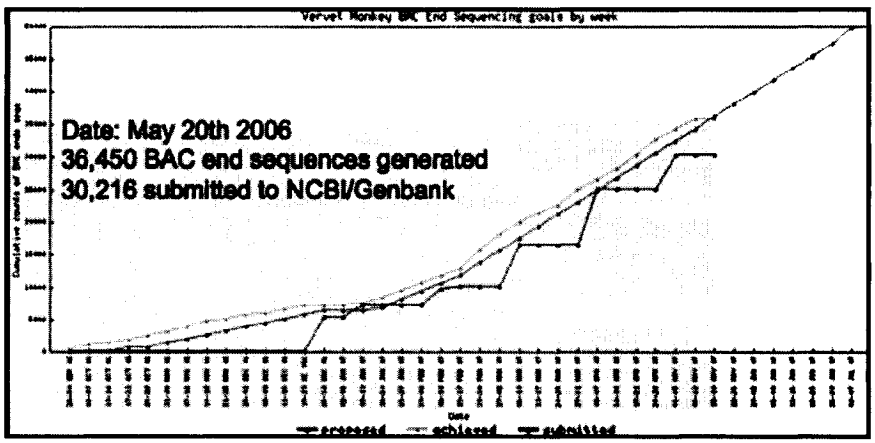

Figure 3: Vervet monkey BAC end sequencing goals by week

Figure has been adapted from www.genomequebec.mcgill.ca/compgen/vervet_web/home.shtml. proposed (RED), achieved (GREEN) and submitted to NCBI/Genbank (BLUE).

Fold coverage is determined by the formula:

Coverage = (Number of clones x Average vervet clone insert size)/Vervet genome size

= (15, 108 x 160,000 bp) / 3,200,000,000 bp

= 0.75

Since two BAC end sequences are submitted for each clone, the value used for the 'Number of clones' is half the number of vervet BAC end sequences submitted to NCBI Genbank (Figure 3).

1.4.3 Vervet BAC library

The CHORI-252 vervet monkey BAC library was constructed by Michael Nefedov in Pieter de Jong's laboratory at BACPAC Resources Children's Hospital Oakland Research Institute (http://bacpac.chori.org/monkey252.htm). Blood was obtained from an adult male vervet monkey (#1994-021) belonging to the UCLA/VA pedigreed colony. DNA was extracted from agarose embedded white blood cells, partially digested

with the enzyme *EcoR1*, size-fractionated, ligated to the *pTARBAC2.1* vector, and transformed into T1 resistant *DH10B* bacteria. The BAC library contains approximately 200,000 BAC clones of average size 160 kb, representing 10-fold redundant coverage of the vervet monkey genome. It is available as frozen glycerol stock cultures organized in a series of 528 384-well microtitre plates or as a collection of eleven 22x22 cm nylon high-density colony filters ideal for screening by probe hybridization.

1.5 Research Aims

The general objective of my thesis work was to develop methods to identify and characterize regions of genome evolution in the vervet monkey with the hope that this will allow us to observe how the vervet monkey and human genomes differ, and that the knowledge gained from this project will help us better understand the mechanisms that have shaped primate and other genomes over time. Under this rather broad umbrella, the following specific questions are addressed in this dissertation.

1. Does the vervet genome differ at the karyotypic level from that of the human and the rhesus?
    - What is the direction of this change?
    - Which vervet centromeres correspond to evolutionary old and new centromere positions?
2. Can centromeric BAC clones in the vervet be identified using a centromeric DNA marker?
    - How does one define a centromeric marker?

- Does the marker allow for the detection of centromeric BACs using a (1) bioinformatics approach and (2) wet lab approach?
- Can marker characteristics be used to estimate the relative position of a candidate centromeric BAC to the functional centromere?
- Can inferences be drawn about centromeric evolution in the vervet with respect to the human?

3. Does comparative mapping to human provide another tool for the identification of centromeric vervet BAC clones?
    - Is it useful for the delineation of old and new vervet centromeres?
    - Does it provide insights about the organization of these regions?
4. Is vervet BAC end sequencing data (at a 0.75X resolution) useful for the detection of genome rearrangements in the vervet?
    - What kind of rearrangements can be detected?
    - Does it provide insight on genome evolution in the vervet?

This thesis is organized as a series of chapters that address each of these questions in turn, employing the appropriate bioinformatic and laboratory techniques. The rationale and objectives of each study are discussed in the preface to each chapter. The final chapter of this thesis summarizes the entire body of work and draws overall conclusions from it.

Chapter Two

Chromosomal evolution in the vervet

2.1 Preface

Evolution of primate genomes can be studied both at the chromosomal, and at the DNA sequence level (Locke et al. 2003). Karyotypic synteny maps provide powerful tools for the study of primate evolution at the chromosomal level. They assist in the establishment of interspecies homology and rearrangement breakpoints, and in the determination of the mechanisms leading to karyotype change in primate species (Yunis et al. 1980; Yunis and Prakash 1982; Finelli et al. 1999; Muller et al. 1999; Kaessmann et al. 2001; Muller and Wienberg 2001). For example, comparative synteny mapping studies in the gibbon with respect to the human and the orangutan detected an unusually large number of chromosomal fissions and translocations, as well as nine neo-centromere formations (Mrasek et al. 2003). The vervet is similar to the gibbon in that the karyotypes of both lineages have been subjected to a large number of rearrangement events. Comparative synteny analyses in the vervet should thus help delineate the history and mechanisms of vervet-specific chromosomal evolution.

2.2 Analyses of a human/rhesus/vervet synteny map

In order to study large-scale karyotypic changes specific to the vervet lineage, in particular chromosome breaks and emergence of new centromeres, I used published studies of genome wide synteny relationships between the vervet and human (Finelli et al.

1999) and the rhesus and human (Weinberg et al. 1992) to create a human/rhesus/vervet synteny map (Figure 4). My three species karyotype map was constructed by taking each pairwise study and visually aligning the banded chromosomes to create a three-way ultrastructural alignment. Since the vervet and rhesus are postulated to have the same Old World monkey ancestor (Goodman 1999), I used the human as an outgroup species.

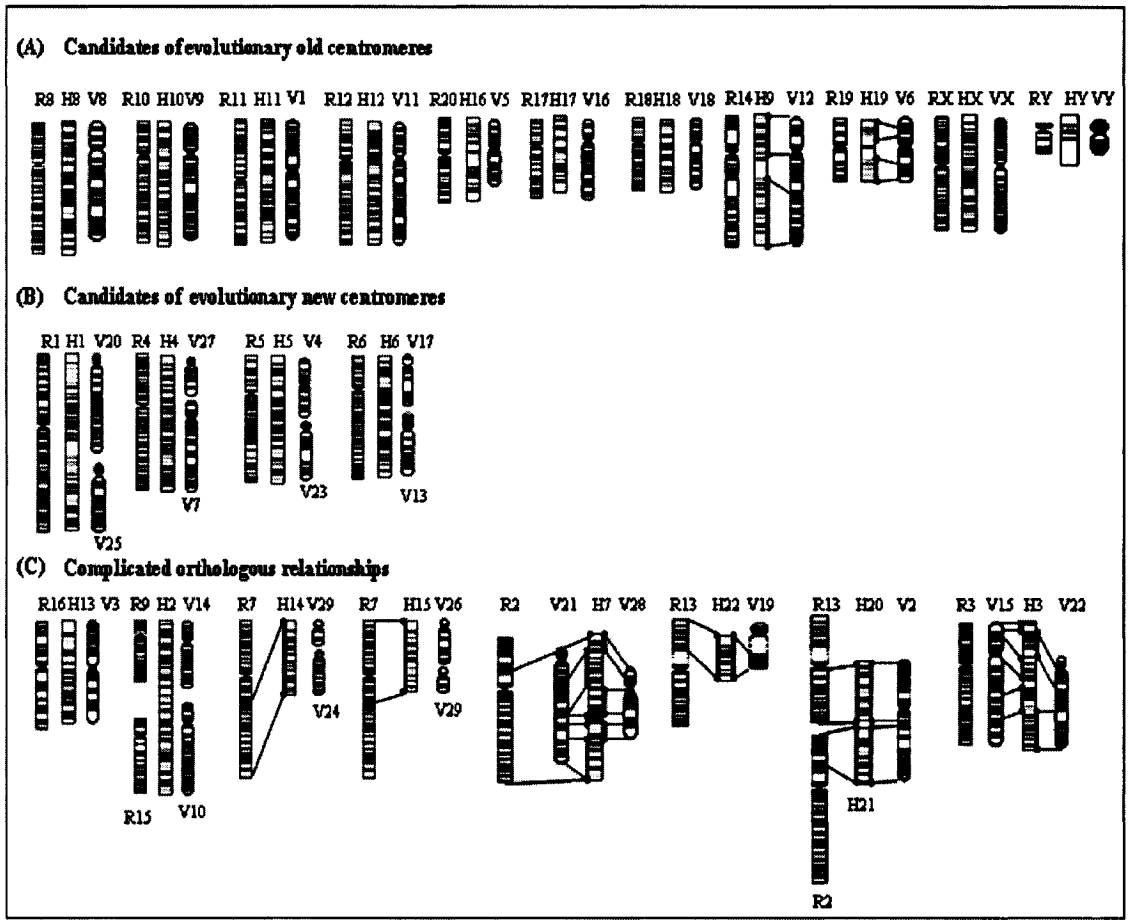

Figure 4: Human, rhesus and vervet chromosomal synteny relationships

This diagram is based on (Finelli et al. 1999) and (Weinberg et al. 1992). Panel A: eleven chromosomes showing evolutionary conserved centromere locations between human, rhesus monkey and vervet monkey. Panel B: four chromosomes showing conserved centromere placement between human and rhesus monkey, but recently diverged chromosome evolution for the vervet monkey. Panel C: thirteen other vervet chromosomes correspond to human and rhesus monkey chromosomes in more complicated relationships.

Observation of synteny in all three species was interpreted as karyotypic features predating (>25 mya) the divergence of the apes and the Old World monkeys.

2.3 Identification of conserved centromere positions

The human genome (n = 23), the rhesus genome (n = 25), and the vervet genome (n = 30) share nine full autosome orthologs, as well as the sex chromosomes (Figure 4, Panel A). As the position of the centromeres appear to have remained unchanged across the three lineages in these eleven chromosomes, they represent candidates of ancestral centromeres (>25 mya) still present in the vervet genome.

2.4 Identification of evolutionary recent centromere positions

I also identified four human chromosomes (1, 4, 5 and 6) showing full orthology between human and the rhesus monkey, but where the vervet monkey now has two chromosomes (Figure 4, Panel B). These eight vervet chromosomes represent evolutionary recent (<9 mya) candidates of chromosome breakage in the vervet. Seven of these newly arisen chromosomes also exhibit evolutionary recent centromere positions (vervet chromosomes 7, 13, 17, 20, 23, 25 and 27). Acquisition of centromere function in 5/7 cases appear to have occurred at completely new positions in comparison to the rhesus and the human, and in 2/7 cases at the chromosomal fission-points.

2.5 Identification of other ultra structural modifications

Twelve vervet chromosomes displayed more intricate orthologous relationships with nine human autosomes and seven rhesus autosomes, making it more difficult to evaluate their centromere history (Figure 4, Panel C). Regardless, I was able to infer that seven of these vervet chromosomes had centromere positions corresponding either to human (vervet chromosomes 26) or to the rhesus monkey (vervet chromosomes 3, 10 and 14) or both (vervet chromosomes 15, 28 and 29). The remaining five vervet chromosomes (2, 19, 21, 22 and 24) exhibited centromere positions that lacked orthology with either human or rhesus synteny maps, and represent additional candidates for novel centromere positions (<9 mya) in the vervet. Emergence of centromere function appeared to have occurred at the fusion-point of two chromosomes in vervet chromosome 2, at a completely new position in vervet chromosomes 19, 21 and 22 and at the fission-point in vervet chromosome 24.

2.6 Mechanisms of lineage-specific karyotype change in the vervet

Finelli et al. (1999) have suggested that the difference in diploid numbers between the vervet (2n = 60) and human (2n = 46) was mainly owing to non-Robertsonian chromosome fissions. This type of fission takes place when a chromosome breaks at a non-centromeric location. My analyses of the human/rhesus/vervet synteny map established that an increased number of fissioned chromosomes distinguishes the vervet karyotype from that of both the human and the rhesus. In addition, the tri-species analyses confirmed that most chromosome breakpoints in the vervet lie outside of the centromere

regions (of both human and rhesus). As expected, the increase in chromosome number in the vervet is also associated with an increase in the number of centromeres. Compared to the 24 centromeres present in the human and rhesus karyotypes, the vervet karyotype exhibits a total of 31 centromeres. I found that approximately a third (at least 12) of the vervet centromeres inhabited evolutionary new positions. In the majority of cases (8/12), acquisition of centromere functions appeared to have occurred at completely new positions relative to the human and the rhesus karyotype. In the remaining cases, emergence of the new centromere position was either at a chromosomal fission (3/12) or a chromosomal fusion (1/12) point.

2.6.1 Centromere repositioning

The current literature advocates that centromere emergence and/or repositioning, together with other chromosomal rearrangements, are the two main promoters of lineage-specific changes in karyotype. In this section and the next, I provide a description of the two mechanisms and how they might have had precipitated vervet-specific karyotype changes.

Evolutionary recent acquisition of centromere function in the vervet may have resulted as a consequence of latent centromere reactivation, centromere spreading, or chromosome fission. In human, reactivation of an ancestral centromere at chromosomal region 15q24-26 has been demonstrated to give rise to ectopic centromere formation. Reactivation at latent centromeres can occur either spontaneously or following rearrangements and does not result in alteration of intervening marker order (Montefalcone et al. 1999; Ventura et al. 2001; Ventura et al. 2003). The endogenous

centromere either undergoes decay or remains active. Retention of activity at the endogenous centromere can result in the emergence of two new acentric chromosome fragments, each with the ability to segregate autonomously (Choo 1997; Hall et al. 2004).

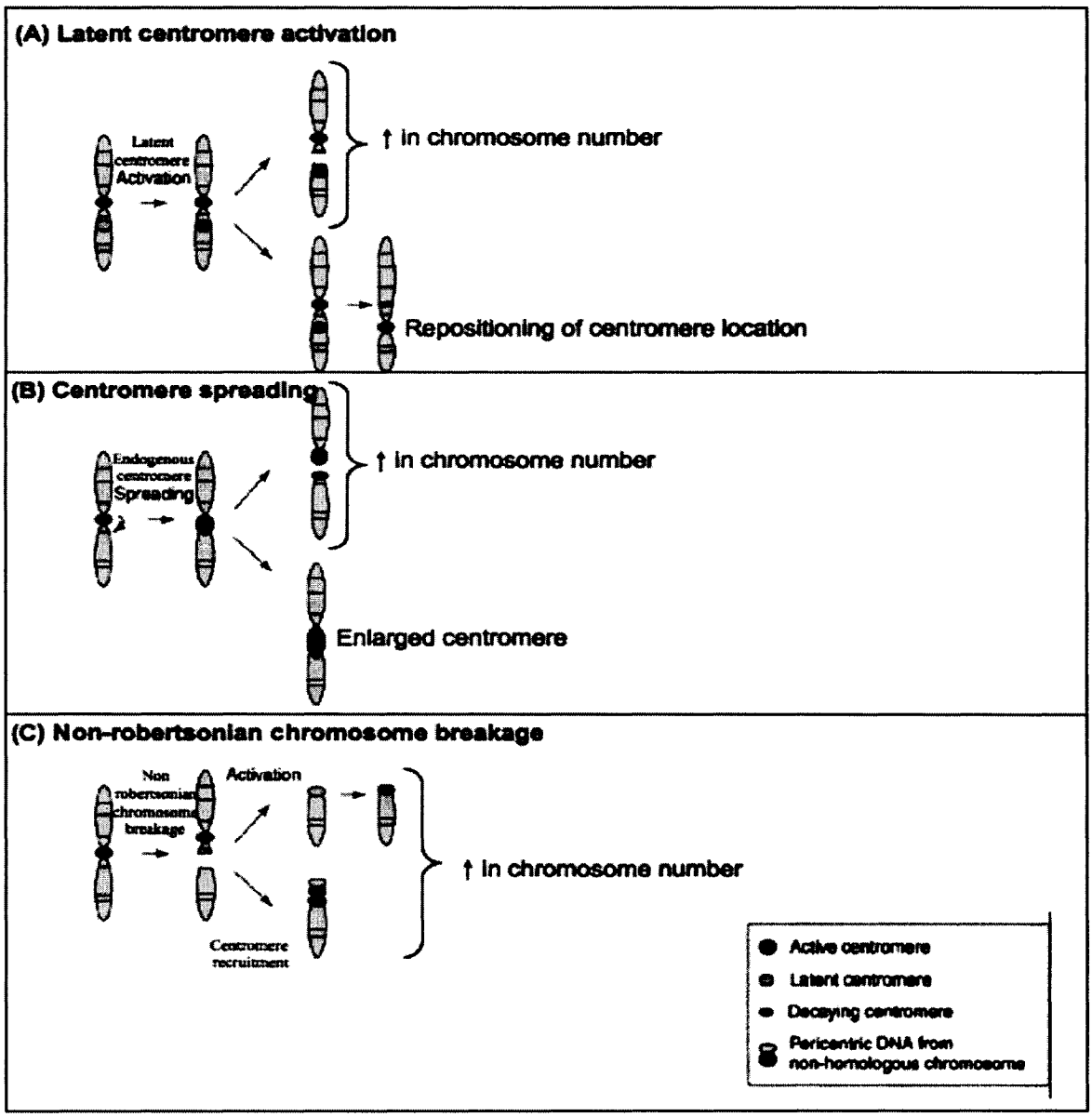

Figure 5: Acquisition of centromere function as a mechanism of karyotype change

This figure has been adapted from (Hall et al. 2004). (A) Latent centromere activation can result in centromere repositioning. If the original centromere also remains active, chromosome breakage can occur and may result in chromosome number increase. (B) Centromere spreading can also lead to chromosome breakage followed by an increase in chromosome number. (C) Chromosome fission events can activate a latent centromere or can be repaired with ectopic centromeric or pericentromeric DNA from non-homologous chromosomes.

In the vervet, latent centromere reactivation (Figure 5, Panel A) may provide a plausible explanation for the centromere number increase and the non-Robertsonian increase in chromosome number.

Gain of centromere function by non-centromeric regions can also occur as a result of centromere spreading. Centromere protein components have been known to spread to linked chromosomal regions and form a larger centromere. This phenomenon has been well documented in *Drosophila* (Maggert and Karpen 2001; Hall et al. 2004). Breakage of the larger centromere in this model can lead to the formation of two separate chromosomes, each with its own centromere (Figure 5, Panel B). Visual examination of the human/rhesus/vervet synteny map did not provide any evidence of centromere spreading-induced chromosome breakage nor centromere function acquisition in the immediate history of vervet karyotype evolution. However, since the karyotypic phenotype produced by such a process might be masked by more recent changes in the vervet, additional in-depth studies are required.

It is most likely that in the vervet ectopic centromere activation was triggered by chromosome breakage (Figure 5, Panel C). Loss of the endogenous centromere (in at least one of the broken fragments) can trigger the activation of a latent centromere. Alternatively, repair of the broken chromosome ends with centromeric DNA from non-homologous chromosomes can result in the emergence of new centromeres. Mouse chromosomes 5 and 6 appear to be chromosome fission products, repaired with fragments of centromeric DNA from other chromosomes (Thomas et al. 2003; Hall et al. 2004). Literature suggests that approximately 20 percent of chromosome fission breakpoints are reused during mammalian evolution, often as sites of new centromere emergence

(Murphy et al. 2005). Intriguingly, the majority of neo-centromeres analyzed to date have been associated with genome rearrangements (Amor and Choo 2002; Ventura et al. 2003).

2.6.2 Genome rearrangements

The major driving force for chromosome evolution in mammals is believed to be genome rearrangements rather than single nucleotide mutations/polymorphisms (Stankiewicz and Lupski 2002; Locke et al. 2003; Stankiewicz et al. 2004). There exist two primary mechanisms of genomic rearrangements, non-allelic homologous recombination and non-homologous end joining. Recent studies show that regardless of the recombination mechanism used, chromosomal rearrangement breakpoints tend to predominate in regions of complex genomic architecture, in particular segmental duplications. Segmental duplications are a class of repetitive element >10 kb in length that share a high level of sequence identity (>90%), and have been hypothesized to predispose regions containing them to recurrent rearrangement by non-allelic homologous recombination (Emanuel and Shaikh 2001; Bailey et al. 2003; Sharp et al. 2005).

Comparative genomic studies to date demonstrate a strong correlation between segmental duplications and lineage-specific changes in karyotype (Bailey et al. 2001b; Eichler 2001; Bailey et al. 2002). Comparisons between the genomes of human and mouse show an enrichment of segmental duplications in regions of chromosome synteny breakpoints (Armengol et al. 2003; Pevzner and Tesler 2003b; Pevzner and Tesler 2003a; Bailey et al. 2004). Genome-wide studies of human and chimpanzee also point to large segmental duplications being responsible "for almost all of the most extreme differences

in chimp chromosome structure, including the emergence of African great ape sub-terminal heterochromatin" (Cheng et al. 2005). Segmental duplications are also thought to be responsible for the rapid lineage-specific large-scale chromosomal rearrangements in gibbon and the squirrel monkey (http://eichlerlab.gs.washington.edu/primate.html). Based on these findings, it is very likely that a sudden increase in segmental duplication rate might have predisposed the vervet genome to undergo extensive chromosome rearrangement events, in particular chromosome fissions and neo-centromere emergence.

2.7 Summary

In summary, I performed a three species comparative karyotype analyses to identify 15 evolutionary old centromeric locations in the vervet genome with respect to the human lineage. Using the method I also delineated 10 centromeric positions that appear specific to the vervet lineage and are candidates of evolutionary recent centromere positions. In addition, I have provided a general explanation of how vervet-specific karyotype changes might have occurred.

Chapter Three

Use of alpha-satellite repeat as a centromere marker

3.1 Preface

In most eukaryotes, the centromere is a specialized domain that mediates critical mitotic and meiotic functions, including kinetochore nucleation, spindle attachment, and sister chromatid cohesion (Henikoff and Malik 2002; Hall et al. 2004; Henikoff and Dalal 2005).

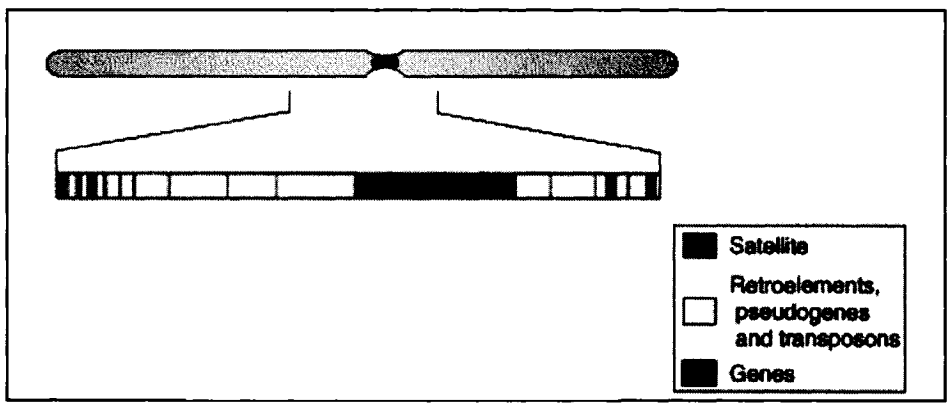

Figure 6: The higher eukaryotic centromere

The figure is adapted from (Hall et al. 2004). Cytologically, centromeres are chromosomal constrictions (green) that nucleate the kinetochore. Centromeric regions encompass large arrays of satellite DNA (red) flanked by pericentromeric DNA (yellow). At the edge of the centromere, pericentromeric DNA gradually transitions into the gene-rich chromosome arms (purple/aqua).

Centromeres of higher eukaryotes contain several megabases of densely methylated, highly repetitive, heterochromatic DNA (Henikoff 2002). The innermost centromeric region, composed of tandemly repeated satellite DNA (Figure 6), has been hypothesised to be the functional domain of the centromere structure. The DNA immediately flanking

the functional centromere is referred to as the pericentromere (Figure 6). Primarily enriched in interspersed repetitive elements and segmental duplications, the pericentromere may also contain domains that contribute to centromere activity (Hall et al. 2004).

The precise centromeric boundaries have been known to vary both between the chromosomes constituting the genome of a single species and between species. Furthermore, studies indicate that though the function of centromeres is conserved throughout eukaryotic biology, their DNA sequences are not (Eichler 1999; Henikoff and Dalal 2005). Absence of sequence homology among the centromeres of distantly related species points to an extremely rapid rate of centromeric DNA evolution (Hall et al. 2004). Primate centromeric regions are rich in genome rearrangements. To date, analyses (phylogenetic and comparative) of pericentromeric regions in human "reveals a mosaic of duplicated and transposed segments of complex evolutionary origin" (Horvath et al. 2000b). This predisposition for ongoing rearrangements makes primate centromeric locations and in particular the vervet centromeric region fertile grounds for the evolution and discovery of lineage specific novel gene functions.

Having identified a set of ancestral and new centromeres in the vervet using comparative synteny analyses, my subsequent goal was to develop procedures for the detection of the corresponding vervet BAC clones. Identification of centromeric clones will facilitate the characterization of evolutionary old and new vervet centromeres, and shed light on centromere evolution in the vervet with respect to the human and other primates. Presented in this chapter and Chapter 4 are two methodologies that I developed for the identification of vervet BAC clones embedded in centromeric regions.

In this chapter, I make use of primate centromeric DNA sequence information to single out centromere associated BAC clones in the vervet. Primate centromeres harbour an impressive array of highly repetitive DNA that is composed primarily of interspersed and tandem repeat elements and segmental duplications (Henikoff and Malik 2002; Rudd and Willard 2004; Schueler et al. 2005). While both interspersed and tandem repeat elements display high copy numbers, they vary based on sequence characteristics and mode of propagation. Interspersed repeats propagate themselves via RNA mediated transposition and consist of sequences scattered throughout the genome. "Tandemly repeated DNA on the other hand expands or contracts by unequal crossing-over or replication slippage" (Horvath et al. 2000a). Commonly known as "satellite DNAs", tandemly repetitive sequences can be either of the satellite, minisatellites or microsatellites variety (Charlesworth et al. 1994). The three subgroups differ in lengths, properties and genomic distributions. Of these, only the satellite class of sequences have been shown to cluster in the heterochromatic rich centromeric and telomeric regions of primate chromosomes (Slamovits and Rossi 2002). These sequences thus might be helpful in the identification of vervet BACs within centromeric regions.

3.2 Human and chimpanzee analysis of centromere markers

In order to recognize vervet BAC end sequences associated with centromeric regions, I started by analyzing the human and chimpanzee genome assemblies for satellite sequences that were distinct to centromeres and present at all centromeres. The UCSC Table Browser was used for this purpose since it allows for the retrieval of a specific subset of records from a track or positional table in a selected genome assembly

(http://genome.ucsc.edu/goldenPath/help/hgTablesHelp.html). In particular, the Table Browser filter option was used as it provided me with the means to restrict the query output to records pertaining to the repeat class 'Satellite' in both genomes.

The data mining results revealed that of the 24 satellite families present in the human genome (*hg17; May 2004 assembly, Build 35; http://genome.ucsc.edu/*) six were associated with centromeres (Table 3). These centromere associated satellite families are alpha-satellite (ALR/Alpha), gamma-satellite (GSAT), gamma-satellite 2 (GSATII), gamma-satellite X (GSATX), human-satellite 4 (HSAT4) and SST1.

Repeat Name	Repeat Class	Repeat Family
ACRO1	Satellite	acrocentric
ALR/Alpha	**Satellite**	**centromeric**
GSAT	**Satellite**	**centromeric**
GSATII	**Satellite**	**centromeric**
GSATX	**Satellite**	**centromeric**
HSAT4	**Satellite**	**centromeric**
SST1	**Satellite**	**centromeric**
(CATTC)n	Satellite	Satellite
(GAATG)n	Satellite	Satellite
BSR/Beta	Satellite	Satellite
CER	Satellite	Satellite
D20S16	Satellite	Satellite
HSAT5	Satellite	Satellite
HSAT6	Satellite	Satellite
HSATI	Satellite	Satellite
HSATII	Satellite	Satellite
LSAU	Satellite	Satellite
MSR1	Satellite	Satellite
SAR	Satellite	Satellite
SATR1	Satellite	Satellite
SATR2	Satellite	Satellite
SUBTEL_sa	Satellite	Satellite
REP522	Satellite	telomeric
TAR1	Satellite	telomeric

Table 3: Satellite families present in the human (*hg17*) and chimpanzee (*panTro1*) genomes

The presence of the same six satellite families in the chimp centromeres (*panTro1; November 2003 assembly, Build 35; http://genome.ucsc.edu/*) established that the six centromeric satellites were not unique to human lineage and thus could be present in additional nonhuman primate species, including the vervet monkey.

Of the six satellite families present in centromeric regions, I found that alpha-satellite was the only one present at all centromeres in both human and chimp. The alpha-satellite DNA is composed of AT-rich monomers that are approximately 171 base pairs in length and organized in a head to tail pattern. To date, alpha-satellite DNA has been detected in all primate species studied, and appears to be unique to the primate lineage (Willard, HF.; 1990). In human, long uninterrupted stretches (3-5 Mb) of alpha-satellite DNA is known to be present at and near functional centromeres (Schueler et al. 2001; Schmutz et al. 2004; Ross et al. 2005; Schueler et al. 2005; Nusbaum et al. 2006; Rudd et al. 2006).

My analysis of alpha-satellite distribution in the human genome (*hg17; May 2004 assembly, Build 35; http://genome.ucsc.edu/*) showed the presence of 155 alpha-satellite regions totalling > 6.9 Mb of sequence (Figure 7). Presence of large alpha-satellite regions (>50 kb) adjacent to the centromere gaps in 18 chromosomes confirmed the enrichment of these repeats near centromeres. Absence of long alpha-satellite stretches in the remaining six chromosomes (4, 13, 15, 17, 18 and 22) showcases the incompleteness of the genome assembly close to centromeric regions. I also detected smaller regions (<1 kb and 1-50 kb) of alpha satellite sequence in 122 non-centromeric locations (Figure 7).

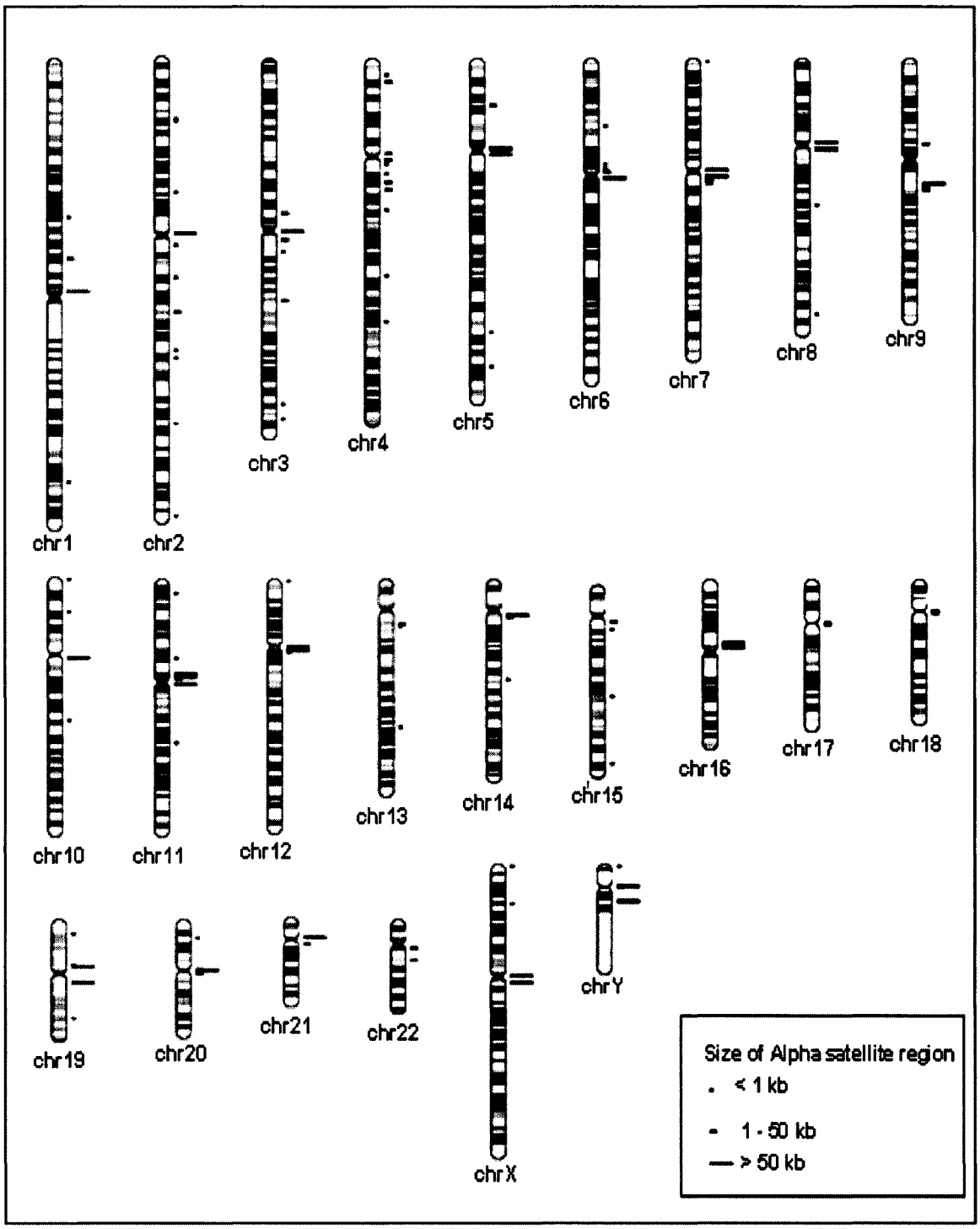

Figure 7: Distribution and frequency of human alpha-satellite repeats

The *hg17* human genome assembly (May 2004 assembly, Build 35; http://genome.ucsc.edu/) was used for this analyses.. All alpha satellite regions have > 60% alpha satellite sequence content. The human chromosome ideograms are based on Furey and Haussler (2003).

3.3 Detection of BACs with ALR content using sequence analysis

In a preliminary study (May 2005) involving 2,755 vervet monkey BAC end-sequences I identified 75 (2.7%) with alpha-satellite (ALR) content. Of the 75 ALR positive ends detected, 60 were the matched pairs from 30 BAC clones, while 15 were observed in a single end of 15 clones. Thus, in total I detected 45 BACs with ALR content. Sequence analysis was performed using RepeatMasker (www.repeatmasker.org) on every BAC end read, followed by retrieval of the output files masked by ALR consensus sequences. RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. When compared to the four primate alpha-satellite consensus sequences present in the RepeatMasker library, all 75 vervet sequences aligned with the highest score to the type *ALRa#*. The sequence similarity between human *ALRa#* and vervet alpha-satellite sequences was > 95 percent.

This study was subsequently expanded to analyze a much larger data set. Using the 18,251 BAC end-sequences available as of March 2006, I identified 761 BAC end sequences from 419 vervet BACs with ALR content by RepeatMasker. Of these, 342 clones (82%) display ALR sequences at both ends. The remaining 77 (18%) clones contain ALR content on one end.

The results of this study indicate that vervet clones containing ALR markers could be successfully identified from BAC end sequences. The identification of clones containing ALR in both ends or one end also indicated that BACs at various locations along the gradient of centromeric ALR content were being discovered. Following alignments of the ALR vervet sequences to the human genome assembly, vervet reads

were localized to centromeric locations, but could not be preferentially associated with any particular centromeres. Other approaches thus needed to be developed to estimate the position of centromeric vervet BACs relative to the functional centromere.

3.4 Detection of BACs with ALR content by hybridization screening

I designed a pair of overgo probe sets (Ross et al., 1999) based on the RepeatMasker alpha-satellite consensus sequence *ALRa#* as it was found to be most conserved between primates and our preliminary vervet monkey sequence data. In tandem, the pair of probes were expected to match 87% (39 of 45) BACs with 4 or less substitutions, and not expected to match any other vervet sequence at this level

The probes were labelled with α32P-dATP and α32P-dCTP and hybridized overnight to a vervet monkey BAC filter using our standard laboratory protocols (Appendix I). The filter was then washed to remove unlabelled probe, and exposed to autoradiography film at -80°C for 2 hours. Hybridizing clones were identified using an in-house program calculating clone addresses using a combination of filter position and signal orientation.

Of the 18,432 BAC clones present on the filter (representing ~0.9 fold coverage of the vervet monkey genome), hybridization signals were detected for 718 (3.9%) clones (Figure 8). Of the 45 clones present on the membrane that had ALRa# content in their BAC end sequences, signal was detected in 36/39 (~92%) BACs computationally predicted to bind with the probes.

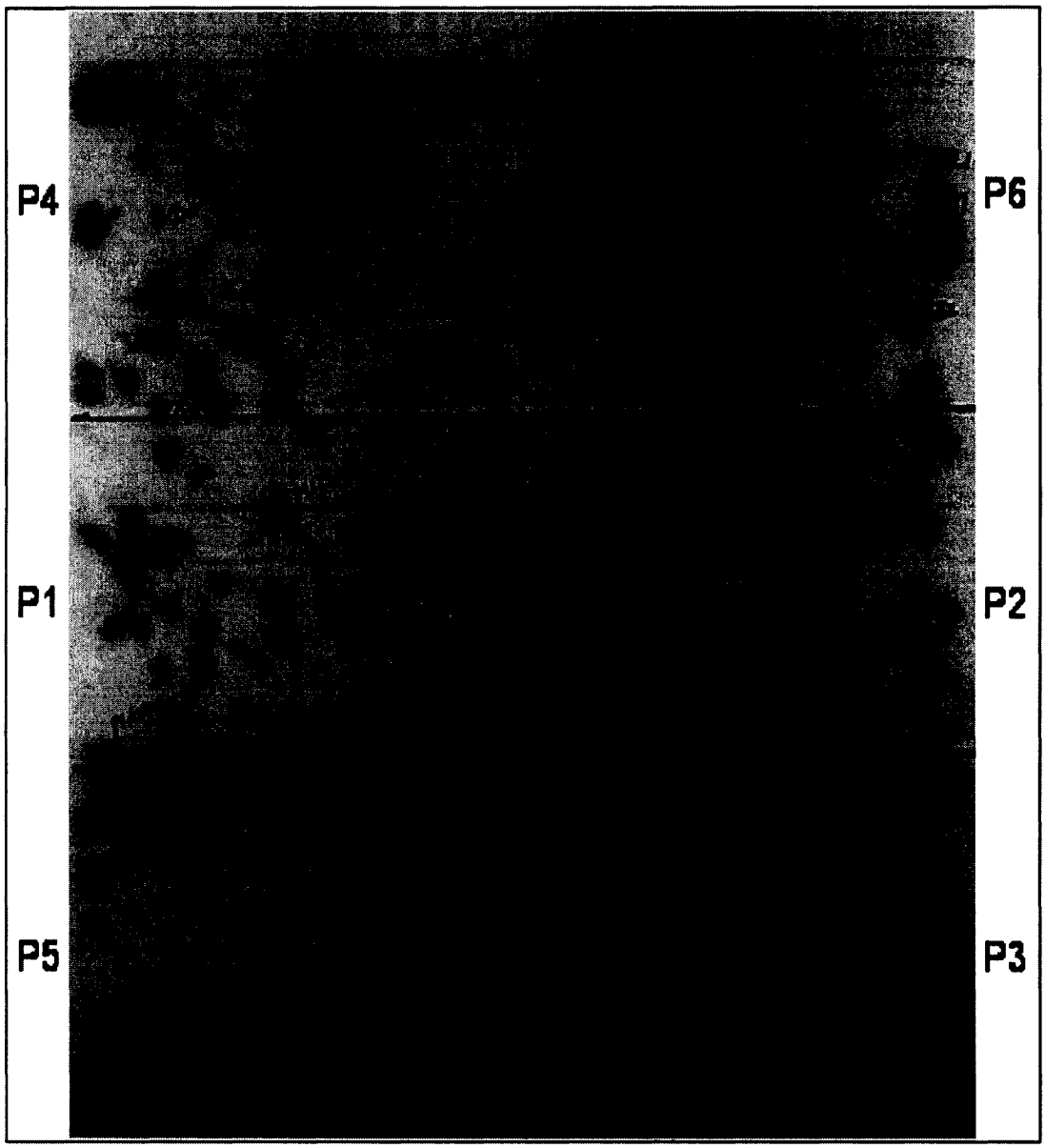

Figure 8: Hybridization screening results

CHORI-252 Filter 1 showing two 36 base pair overgo probes (Probe 1 and Probe 2) hybride to 718 vervet BACs containing alpha satellite repeat sequence.
Overgo Probe 1: ATATTTGGAAGCCCATAGAGGGCTATGGTGAAAAAG
Overgo Probe 2: TATGGTGAAAAAGGAAATATCTTCAGATAAAAACTG
The CHORI-252 vervet monkey BAC library is gridded onto eleven 22x22 cm nylon high-density filters for screening by probe hybridization. Each hybridization membrane represents 18,432 (48 plates, each plate 384 wells) distinct vervet monkey BAC clones, stamped in duplicate.

The *ALRa#* hybridizing BACs were randomly distributed across the filter, with an average of 15 alpha-satellite sequence containing BACs per 384-well plate (Figure 9).

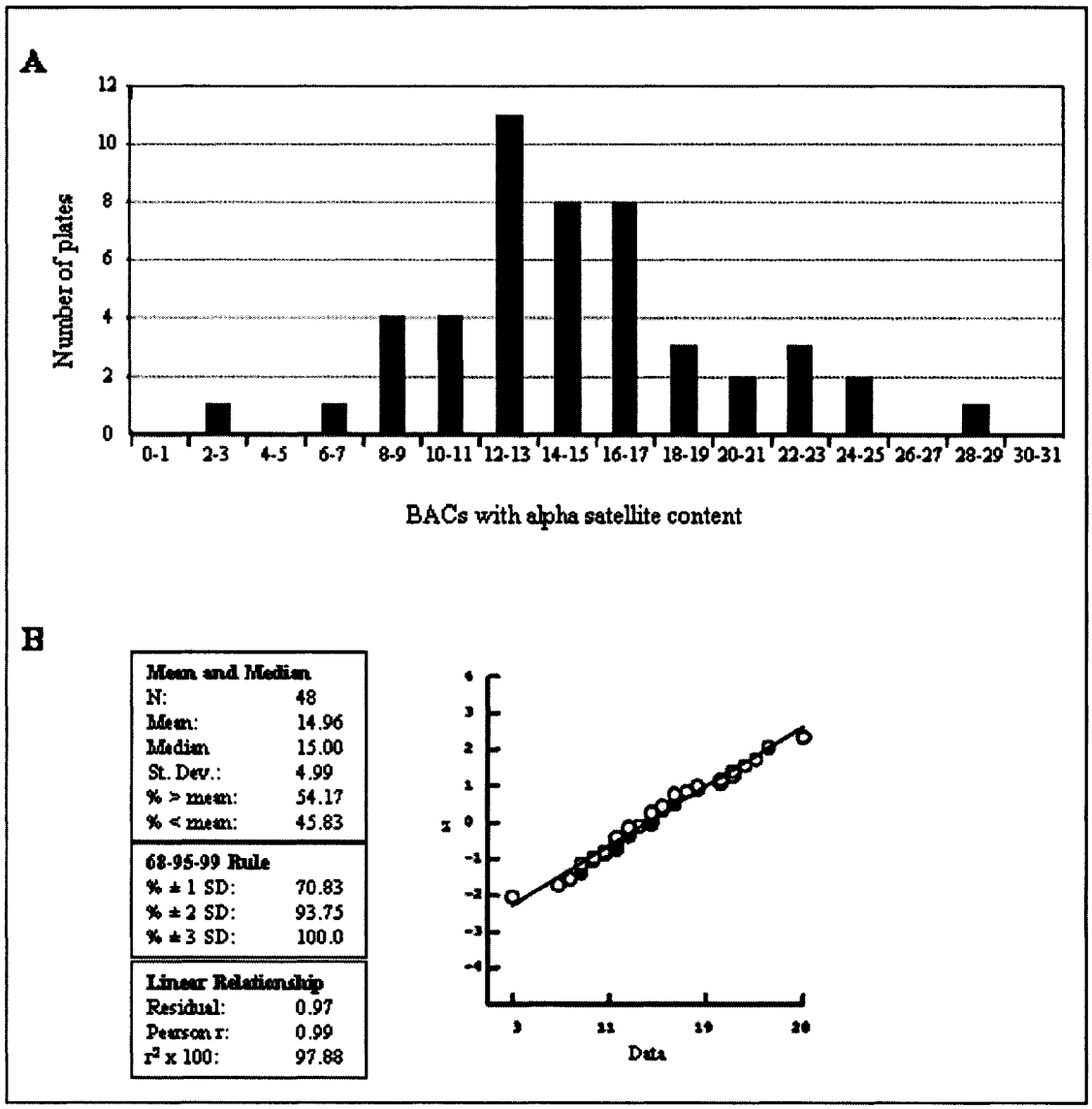

Figure 9: Alpha-satellite content in the CHORI-252 Filter 1

(A) Distribution of the number of plates with varying frequencies of alpha-satellite containing BACs. (B) Statistical analyses showing that the alpha-satellite containing BACs in CHORI-252 Filter 1 follows a normal distribution.

The literature suggests that ~2% of the human genome is composed of ALR sequences (Guy et al. 2003; Eichler and Frazer 2004; Rudd and Willard 2004), which is in general agreement with my analysis of the ALR content and distribution in the human genome assembly. As the vervet monkey genome is comparable in overall size (~3 Gb)

yet contains more centromeres, I believe that the higher proportion of vervet clones containing ALR I observed (~4%) is reasonable.

3.5 Reproducibility of ALR BAC detection by the two techniques

Our group has already end-sequenced 38 plates (plates 11-48) of the 48 CHORI 384-plates present on hybridization filter 1.

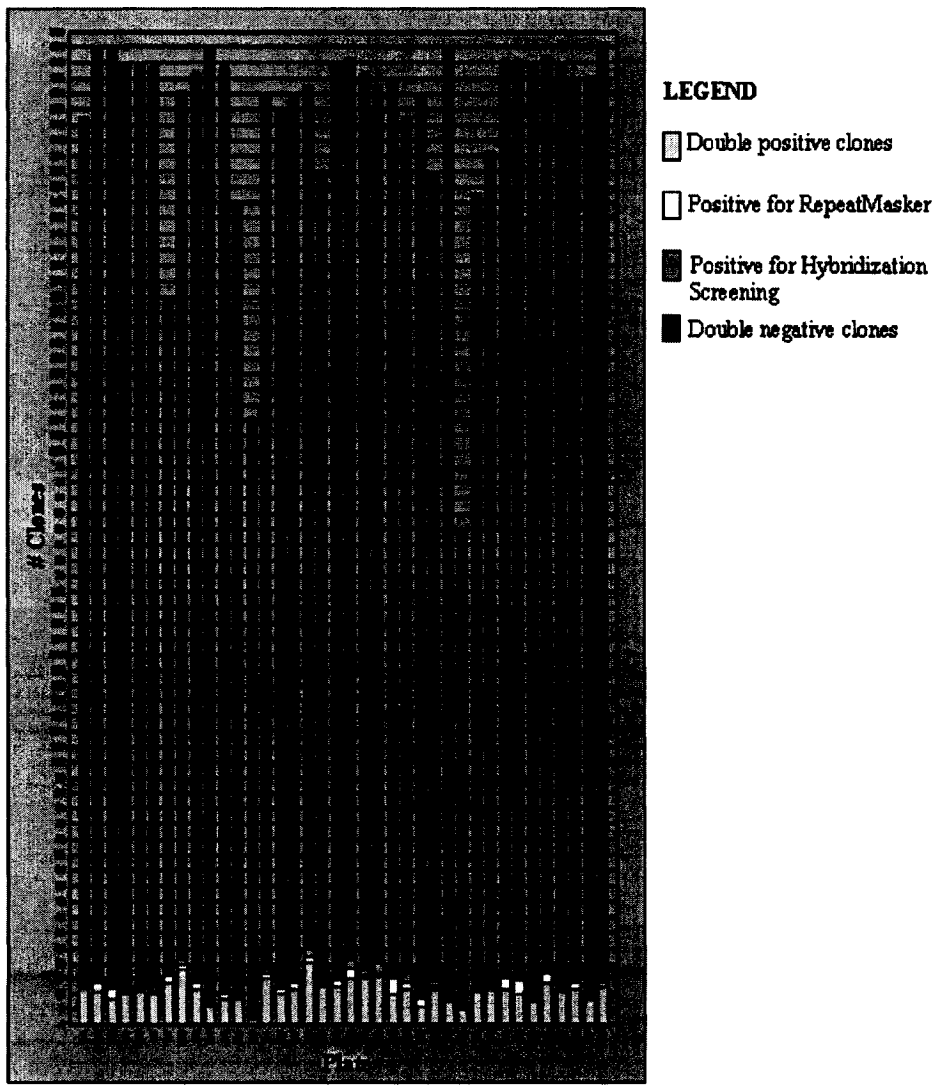

Figure 10: Reproducibility of vervet alpha-satellite BAC detection

Count per plate of vervet alpha satellite containing BACs detected by either RepeatMasker ot Hybridization Screening, or both.

I found that of the 13, 847 vervet BACs with two quality ends, alpha-satellite content was detected in a total of 626 by either RepeatMasker or hybridization screening methods. Since 516 of these BACs were detected to contain alpha-satellite sequences by both methods, I calculated an overall concordance rate of 83% between the two methods (Figure 10).

Aside from the value of repeat masking and hybridization screening for the identification of centromere-associated BACs, our group has also used the data generated by the two methods as a quality assessment of plate naming accuracy in our sequence platform pipeline. By analyzing the reproducibility of alpha-satellite detection by the two methods on a per plate basis, we have shown that the production-sequencing pipeline has not been subject to errors in plate naming. This assessment, combined with other assessments of sequence quality and end-pairing accuracy, has enabled us to document the overall quality of the BAC end sequence resource being generated at our center and submitted to NCBI/Genbank.

3.6 Presence of divergent ALR monomers in the vervet genome

It has been reported that alpha-satellite monomers present in the vervet chromosomes are of the primordial (*S1/AGM*) type (Alexandrov et al. 2001). This primordial alpha-satellite is suggested to be efficiently homogenized throughout the vervet whole genome and nearly identical in all chromosomes. Contrastingly, the great apes and some Old World monkey lineages (for example rhesus) are believed to possess more recently diverged (*S2*, *R1*, *R2*, *M1*, *J1*, *J2*, *D1*, *D2* and *W1-5*) alpha-satellite monomers (Alexandrov et al. 2001; Rudd et al. 2006).

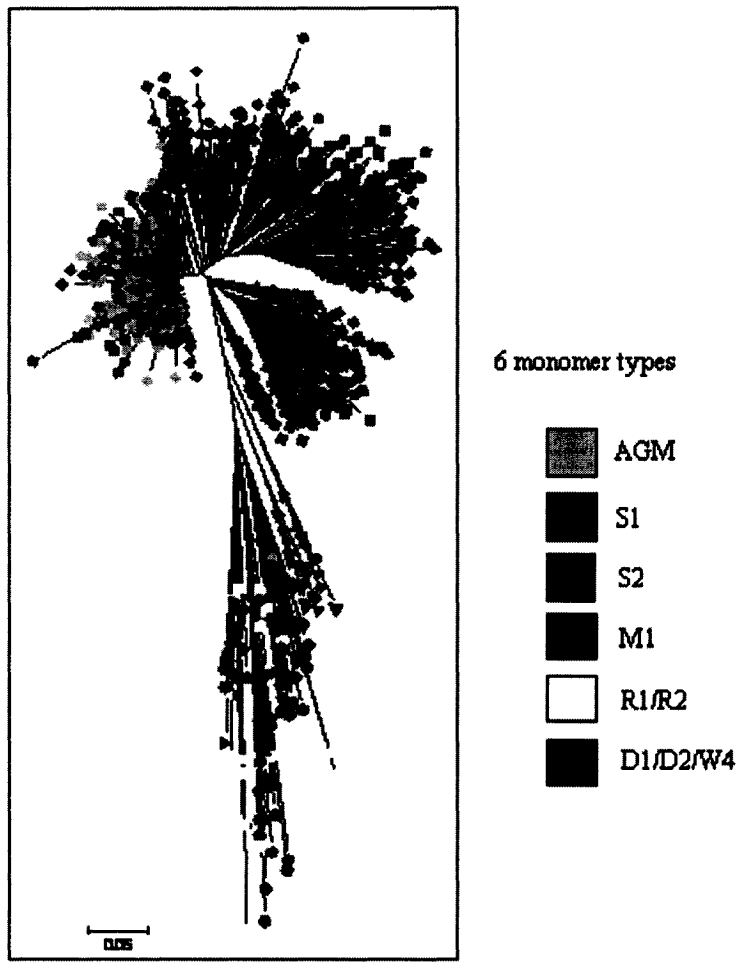

Figure 11: Alpha-satellite monomer types detected in the vervet

Phylogenetic tree was generated by initially aligning the 1,760 alpha-satellite monomers using the program CLUSTALW (http://www.ebi.ac.uk/clustalw/). The resulting monomer alignment was analyzed by the program MEGA (http://www.megasoftware.net/) to generate an unrooted neighbour-joining tree. The tree shows that *S2*, *M1*, *R1*, *R2*, *D1*, *D2* and *D4* monomers present in some of vervet monkey BAC end sequences are divergent from the primordial *AGM* and *S1* monomer types.

To assess the monomer types present in the vervet I isolated 1,760 alpha-satellite monomer sequences. Figure 11 shows the six different alpha-satellite monomer types were identified (using a custom RepeatMasker library) in this collection, namely: *AGM*, *S1*, *S2*, *M1*, *R2* and *W4*. The occurrence of approximately 12% monomers as belonging to divergent monomer types (*S2*, *M1*, *R2* and *W4*) in my sampling is contrary to existing published reports (Alexandrov et al. 2001) and indicates that alpha-satellite organization

in the vervet monkey is more complex than originally believed. The difference in results can be accounted for by the fact that my analyses consisted of >1700 alpha-satellite sequences that were obtained using shot gun sequencing, while Alexandrov and colleagues studied only a limited number (18 sequences) of vervet monkey centromeric sequences, a very small sample size that makes the detection of the rarer forms of alpha-satellite monomers difficult. The result presented in my thesis is thus a more accurate reflection of the alpha-satellite monomer types present in the vervet monkey.

3.7 The use of ALR organization to predict centromeric positions

To date, DNA sequence alignments have been unable to uniquely place vervet BAC end sequences with alpha-satellite content to specific human centromeres, thus impeding direct comparison between vervet and human orthologous centromeric regions. It is well established that in the human genome, two different forms of alpha-satellites exist, namely, a monomeric and a higher order form (Rudd and Willard 2004)

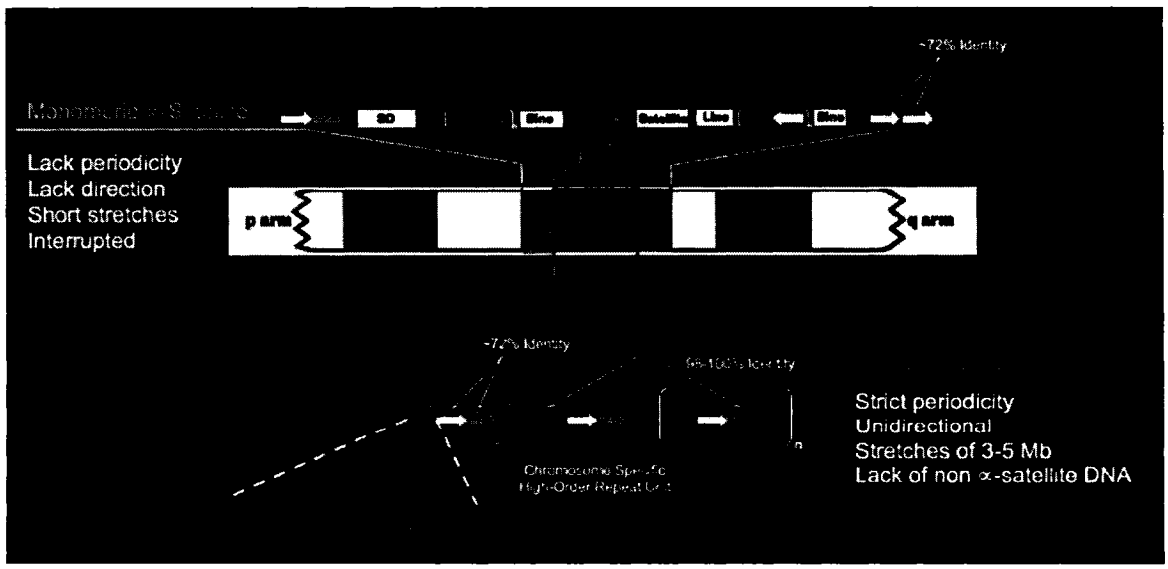

Figure 12: Organization of alpha-satellite repeats at human centromeric regions
This figure has been adapted from (Schueler et al. 2005).

The two forms are spatially separated, with the higher order form being present at or near functional centromeres (Figure 12). Higher order alpha-satellite is composed of chromosome specific groupings of alpha-satellite monomers that are then arranged in a tandem head-to-tail configuration to form an array (Alexandrov et al. 2001). Monomers within a higher order repeat unit have average pairwise sequence identities of ~ 72% while adjacent pairwise repeat units are typically 98-100% identical. Overall, centromeric alpha-satellite arrays tend to be extremely long (3-5 Mb), unidirectional, and display strict periodicity. In contrast, "the long-range structure of pericentromeric alpha satellite regions is disorganized and stratified" (Schueler et al. 2005).

If the vervet monkey genome shows the same trend as the human, then I expect long, organized, uninterrupted arrays of ALR sequences around centromeres, with small islands (<50 kb) interspersed throughout the rest of the genome. Keeping with this, vervet BACs (average insert size ~160 kb, http://bacpac.chori.org/monkey252.htm) within ALR rich regions should display ALR content at both ends. As mentioned previously, I have identified 419 vervet clones with ALR content by repeat masking a total of 18,251 BAC end-sequences. Of these, 342 clones (82%) display ALR sequences at both ends and thus are inferred to occupy centromeric locations in vervet monkey genome. The remaining 77 (18%) clones contain ALR on one end and span either genomic regions of small (< 50kb) ALR islands or inhabit pericentromeric regions distal from the functioning centromere.

I have followed up this general analysis of ALR content with a more refined analysis of ALR higher order structure between and within independent BAC clones. I have been able to identify examples of end sequences sparsely populated with ALR monomers ($\leq$ 2 copies) as well as end sequences almost entirely composed of ALR

monomers (≥ 4 copies). The ALR-rich end sequences tend to be composed of the S1 and AGM classes of ALR monomers. In many cases, an ALR-rich BAC end also has corresponding BAC end that is ALR-rich. In addition, the orientation in the sequences of these BAC clones has been maintained between one end and the other, leading me to hypothesize that the BACs have ALR content throughout. These BACs represent my best candidates of clones embedded in regions in close proximity to the functional centromeres (Figure 13).

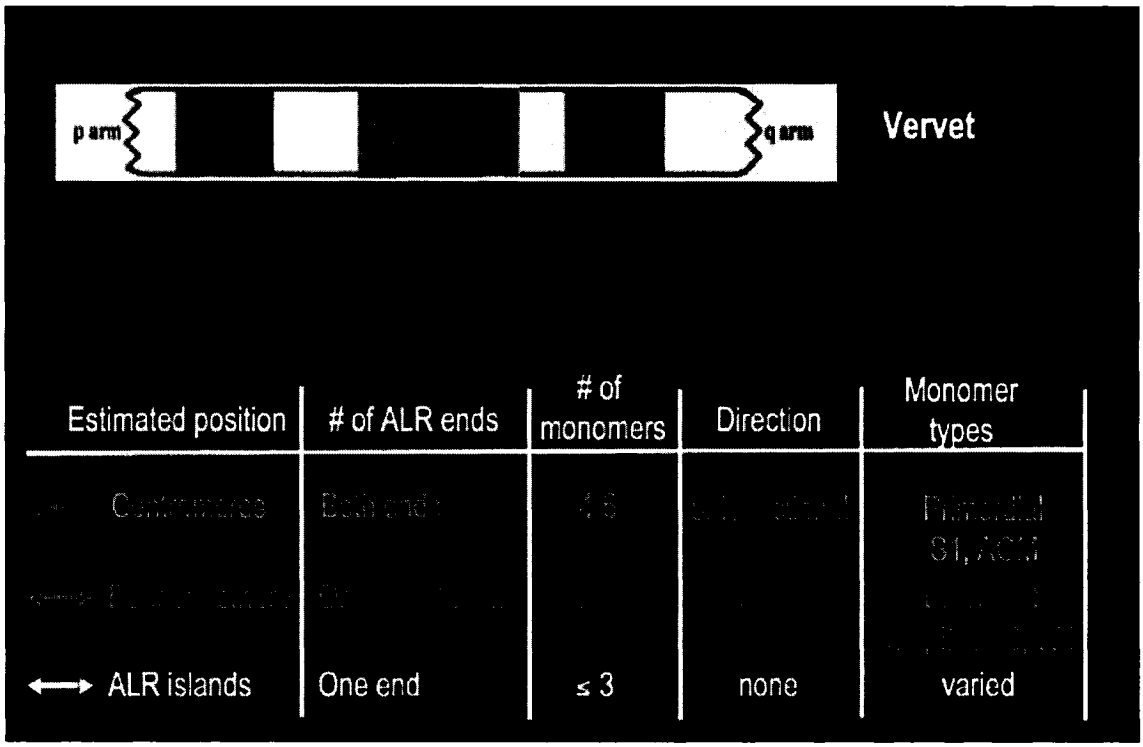

Figure 13: Inferring relative position of vervet BACs by alpha-satellite content

On the other hand, the end sequences containing *M1*, *S2*, *R2* and *W4* type monomers lack complete ALR content and any discernible directionality. I thus predict these BACs are positioned away from the functional centromere and into the pericentromeric regions, or occur at the sporadic small ALR islands similar to those found in human (Figure 13).

3.8 Summary

In the first half of this chapter, I analyzed the human and chimpanzee genome assemblies for sequences that were distinct to, and present at all centromeres. My conclusion from this study was that the alpha-satellite sequence was a robust marker of centromeric DNA and could be used to identify candidate centromeric vervet BAC clones using both sequence analyses and hybridization screening methods.

In the second half of this chapter, my analyses of vervet alpha-satellite monomers led to the identification of alpha-satellite monomer types previously unrecognized in the vervet. In addition, I generated preliminary evidence that the organization of centromere associated alpha-satellite in the vervet differs from that of the human. In human, the DNA sequence at the functional centromere is composed largely of evolutionary recent alpha-satellite monomers. My analyses suggest that in the vervet the opposite might be true, in that the functional centromere is comprised of primordial alpha-satellite monomers while there is an accumulation of evolutionary recent monomers in the pericentromeric regions. Given the few species with genome sequencing completed or underway, it is not possible at this time to determine whether the human/chimp or vervet alpha-satellite centromeric frequency and distribution is most representative of primates in general.

Chapter Four

Chromosome Walking Into Pericentromeric Regions

4.1 Preface

Comparative mapping contributes to the understanding of primate phylogeny and evolution. In this chapter I use a comparative mapping strategy (with respect to the human genome) to parse out vervet BAC clones embedded within centromeric locations. BAC end data obtained from these centromeric/pericentromeric vervet clones should help in the characterization of old and new centromeric locations in the vervet and provide us with insights into the dynamics of centromere evolution.

4.2 Estimation of human pericentromeric boundaries

To demarcate the pericentromeric boundaries of human chromosomes, I used the presence of the six previously established centromeric satellite families (ALR, GSAT, GSATII, GSATX, HSAT4 and SST1) and segmental duplications as guides. Pericentromeric regions in human display an approximately threefold enrichment in segmental duplications compared to the genome average (Zhang et al. 2005). Overall, this approach allowed me to distinguish 39 pericentromeric edge locations belonging to the 22 autosomes (Table 4A). The heterochromatic p arm architecture of the five acrocentric chromosomes (13, 14, 15, 21 and 22) hindered delineation of pericentromeric boundaries in these regions.

Table 4A		
Human Chromosomes	Estimated Locations of Human Pericentromeric Edges	
	p arm	q arm
1	120,500,000	146,748,776
2	84,000,000	133,000,000
3	73,160,900	95,111,649
4	48,926,494	70,500,000
5	45,944,407	50,000,000
6	52,700,000	66,600,000
7	55,300,000	76,000,000
8	43,200,000	48,400,000
9	33,300,000	68,600,000
10	37,000,000	52,300,000
11	48,300,000	55,000,000
12	31,100,000	37,300,000
13	gap	24,700,000
14	gap	20,700,000
15	gap	33,000,000
16	28,200,000	46,300,000
17	14,000,000	23,200,000
18	14,100,000	18,300,000
19	22,000,000	33,100,000
20	23,600,000	34,000,000
21	gap	14,500,000
22	gap	24,500,000

Table 4B	
Distance from Centromere Gap	# of pericentromeric edges
< 1Mb	7
1- 5 Mb	15
5.1 -10 Mb	7
10.1 - 15 Mb	6
15.1 - 20 Mb	7
20.1 - 25 Mb	1
25.1 - 30 Mb	0
30.1 - 35 Mb	0
35.1 - 40 Mb	1

Table 4: Human pericentromeric edges

4A: Estimated locations of the pericentromeric edges for the 22 human autosomes. Note: 'gap' refers to gaps in the human genome (*hg 17*) assembly

4B: Distance from centromere gap and the frequency of pericentromeric edges.

When I calculated the distance of the demarcated pericentromeric edges to their corresponding centromere gaps, I observed they ranged widely from less than 1 Mb to 40 Mb. Table 4B provides a count of pericentromeric edges at various distances from the centromere gaps. The total size of the centromeric/pericentromeric region in the autosomes using the above approach was calculated to be approximately 396 Mb of the assembled human genome.

4.3 Test for cloning bias in human pericentromeric regions

To assess whether pericentromeric regions have any obvious cloning bias that would reduce their presence in a BAC library (and thus hinder the discovery of pericentromeric BACs), one of my objectives was to check whether pericentromeric and euchromatic regions in the human genome displayed differential BAC clone frequencies. I used my predicted set of human pericentromeric regions and matched it with an equivalent number of non-centromeric/non-pericentromeric/non-telomeric regions. My non-pericentromeric control locations in the genome were determined by calculating the mid-point position between the tip of the appropriate p or q arm and my estimated pericentromeric boundary. BAC clones spanning each of the 78 genomic positions (39 pericentromeric boundaries and 39 controls) were obtained utilizing a Perl program that I wrote (Appendix 1). Most clones (~60%) in this sampling were from the RPCI-11 library, a finding consistent with the composition of the current whole-genome BAC map. Build 35 of the human genome (*hg17*) contains a total of 207,973 BAC clones derived from six main BAC libraries covering ~3.1 Gb of the genome. On average, this amounts to 11 BAC clones spanning any given base pair of the assembled genome.

My two groups of pericentromeric and non-pericentromeric positions had an average of 11 and 12 BAC clones respectively. An independent two-tailed student t test found no significant difference in the average number of BACs intersecting pericentromeric edges and euchromatic regions of the genome. Figure 14 shows the variation from the mean (magnitude and direction) of BAC clone frequencies at pericentromeric edges and euchromatic regions.

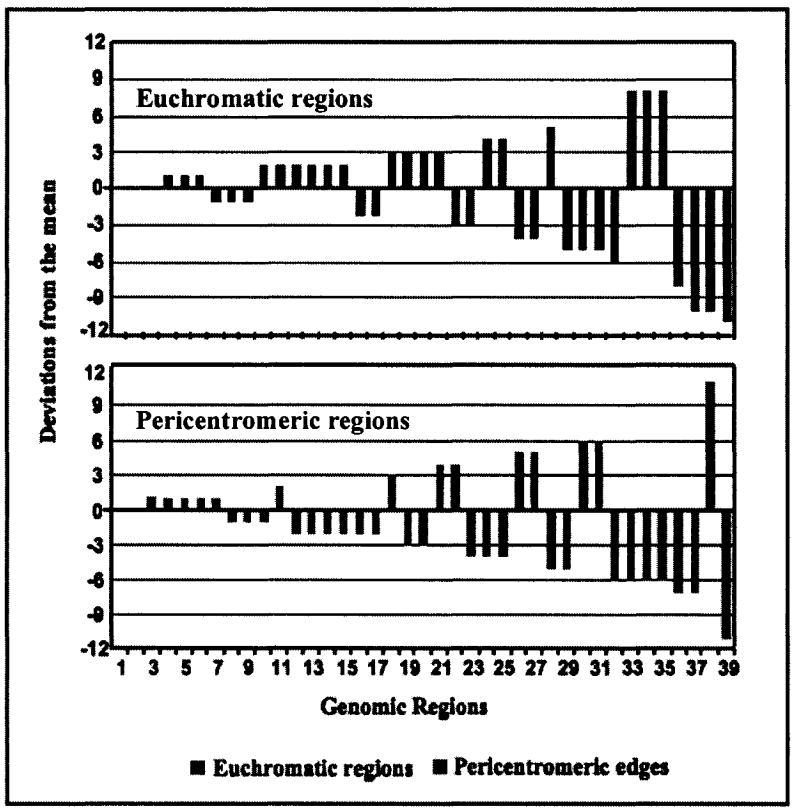

Figure 14: Absence of cloning bias in the human pericentromeric regions

The graps presented show the variations from the mean of BAC clone coverage in the pericentromeric and the euchromatic regions of the human genome assembly (*hg 17*).

Overall, the analysis demonstrates that in the human genome, BAC clone frequencies are evenly distributed and there does not appear to be an under-representation of clones in the pericentromeric regions. Presuming a similar distribution of BAC clones in the vervet monkey BAC library, then there should be an appropriate coverage of pericentromeric edges in the vervet BAC library, which will become recognizable as the BAC end sequencing progresses.

4.4 Identification of pericentromeric vervet BACs

Vervet pericentromeric BACs were identified by subjecting all available quality passed BAC end sequences to human BLAT searches. The BLAT software allows for rapid cross-species alignments of DNA, protein, translated DNA or translated RNA

(http://genome.ucsc.edu). The most probable chromosomal location (chromosome, start, end, sense) of all vervet BAC ends was found by BLAT mapping using gfClient version 32 (minScore = 0 and minIdentity = 0) on gfServer version 29 (type = nucleotide, tileSize = 11, stepSize = 5 and minMatch =0). The protocol used to establish the orthologous location of a vervet BAC end sequence (Sp6 or T7 end) on the human genome entailed the (a) determination of the location in the human genome where the vervet sequence matches best, and (b) verification that no other site in the human genome matches the vervet sequence nearly as well. These requirements were enforced, by using a combination of a minimum threshold score (BLAT score of >125) and a requirement that the "best" score be at least 50 points higher than the "next best" score.

Vervet clone positions with respect to the human genome were interpreted based on the combined orthologous mapping information from each *T7* and *Sp6* pair of reads (*T7* and *Sp6* are forward and reverse primers). Clones were classified as either fully mapped (both ends mapping concordantly to a unique location in the human genome), half-mapped (a single end mapping to a unique location, un-mapped (no end mapping to a unique location), and discordantly mapped (each end mapping to a distinct and separate locus).

Using the above methods, I identified 501 vervet BAC clones mapping to orthologous pericentromeric locations in the human genome. Figure 15 provides a graphical representation of the distribution of these pericentromeric mapping vervet clones per human autosome. Detection of orthologously mapping pericentromeric vervet clones in 21/22 autosomes at relatively low coverage (at most 1.5X) of the vervet genome

also provides evidence of the absence of cloning bias in the distal pericentromeric regions.

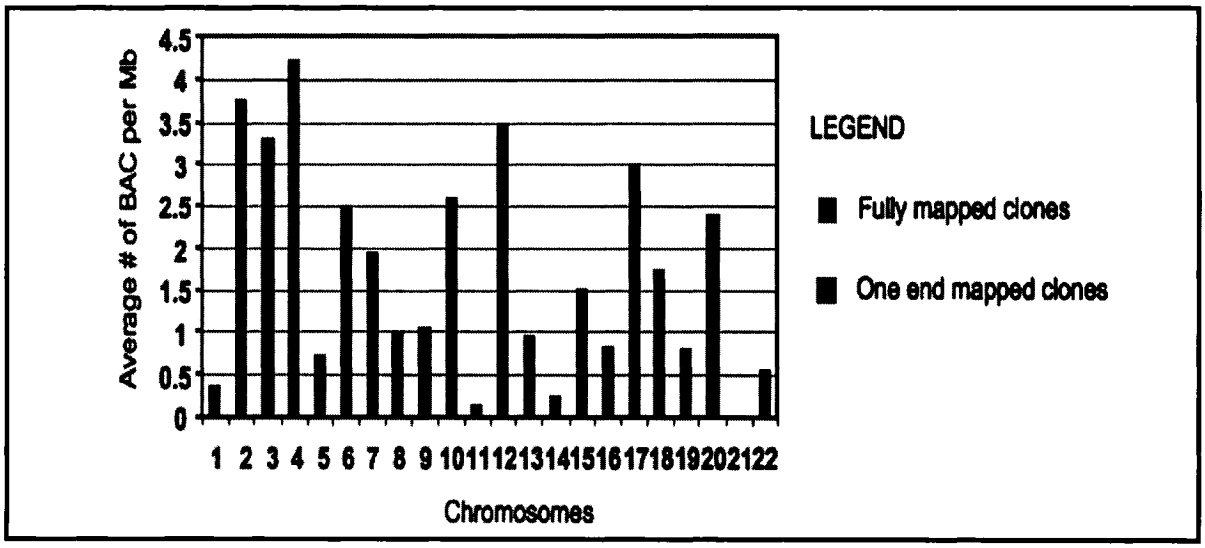

Figure 15: Pericentromeric vervet clones.

Chromosomal distribution of vervet BAC clones mapping to orthologous pericentromeric locations in the human genome. To date, I have detected 501 fully mapped clones in pericentromeric locations and 193 half mapped clones. However, since the sizes of the pericentromeric region vary between chromosomes, the data is best presented as the average number of BACs per megabase. This analysis had a data freeze date of June 1, 2006.

In Table 5, I have provided a list of 29 vervet pericentromeric clones that were detected closest to the centromere gaps. Of these, the majority (25/29, 86.2%) were detected within 5 Mb from the centromere gaps, while the remaining (4/29, 13.8%) were identified at distances ranging from 7.6 Mb – 20.5 Mb. All 29 clones can be used as chromosome arm anchors from which one can begin to analyze the characteristics of deep pericentromeric and centromeric regions in the vervet. BLAT analyses did not detect any pericentromeric mapping vervet BAC for 15 chromosomal arms (Table 5). For these chromosomal arms the BAC clones mapping most proximal to the pericentromeric boundaries can be used as anchors. Such clones could only be detected for 10/15 chromosomal arms (1p, 3q, 4p, 5p, 8p, 11pq, 18p, 19q and 21q).

Chromosome	Distance of pericentromeric edge from centromere gap (Mb)	BAC clone (p arm)	Distance of vervet BAC clone from centromere gap (Mb)
1p	0.7	-	-
1q	23.4	CH252-086B08	20.5
2p	7.7	CH252-097F20	2.7
2q	38.2	CH252-029G21	1.4
3p	17.4	CH252-077N21	0.4
3q	1.6	-	-
4p	0.6	-	-
4q	18.0	CH252-097M02	0.5
5p	0.5	-	-
5q	0.6	CH252-030L10	0.5
6p	6.2	CH252-036N12	1.7
6q	4.7	CH252-066M02	0.4
7p	2.6	CH252-041O22	1.5
7q	15.1	CH252-021J13	2.5
8p	0.8	-	-
8q	1.4	CH252-066M01	0.3
9p	12.7	CH252-019K05	7.6
9q	19.6	CH252-028D07	19.3
10p	2,.2	CH252-061M22	1.4
10q	10.7	CH252-050E03	1.2
11p	3.1	-	-
11q	0.6	-	-
12p	3.6	CH252-039N21	1.1
12q	1.2	CH252-068N04	0.9
13p	n/a	-	-
13q	6.8	CH252-076A16	1.7
14p	n/a	-	-
14q	2.6	CH252-032I09	1.9
15p	n/a	-	-
15q	14.7	CH252-011B22	3.2
16p	6.9	CH252-046L15	3.3
16q	9.4	CH252-041G24	8.1
17p	8.2	CH252-020G17	1.1
17q	0.9	CH252-061P24	0.5
18p	1.3	-	-
18q	1.5	CH252-025G08	0.07
19p	4.9	CH252-035J23	3.2
19q	3.2	-	-
20p	2.7	CH252-097A19	1.4
20q	6.0	CH252-036K18	1.7
21p	n/a	-	-
21q	1.2	-	-
22p	n/a	-	-
22q	10.2	CH252-069C12	1.7

Table 5: Pericentromeric vervet BAC clones closest to centromere gaps

Provided are 29 vervet BAC clones mapping to orthologous pericentromeric locations in the human genome and closest to the centromere gaps. '–' denotes a lack of orthologously mapping vervet BAC clones within the defined pericentromeric boundaries for the indicated human chromosomal arms. This analysis had a data freeze date of June 1, 2006.

In Table 6, I have provided the names of these clones and their distances from the pericentromeric boundaries and centromere gaps.

Chromosome #	Arm	BAC Clone ID	Distance of vervet BAC clone from pericentromeric boundary (Mb)	Distance of vervet BAC clone from centromere gap (Mb)
1	p	CH252-084O20	0.4	1.0
3	q	CH252-097D09	0.06	1.7
4	p	CH252-051N08	0.002	0.6
5	p	CH252-041N04	0.05	0.5
8	p	CH252-082K07	0.1	0.8
11	p	CH252-034M22	0.01	3.2
11	q	CH252-012J22	1.3	1.8
18	p	CH252-065F17	0.4	1.7
19	q	CH252-030J16	0.4	3.6
21	q	CH252-025H18	0.001	1.2

Table 6: Vervet BAC clones most proximal to the pericentromeric boundaries

For the 10 non-acrocentric chromosome arms that lacked pericentromeric clones, I have provided the names of the BAC clones that mapped closest to the pericentromeric boundaries. All clones were located within 2 Mb of their corresponding pericentromeric boundaries and within 4 Mb of their corresponding centromere gap position.

The remaining 5 chromosomal arms (13p, 14p, 15p, 21p and 22p) embody the knobs of acentric chromosomes and represent unassembled regions in the current human genome assembly. BLAT analysis was thus unable to detect any orthologously mapping vervet clones to these regions.

In addition to identifying fully mapped vervet clones orthologous to human pericentromeric region, I also identified 193 clones where one end maps unambiguously to a pericentromeric region, while the other end, despite being of quality sequence, fails mapping (Figure 15). Sequence analyses (using RepeatMasker) revealed that 192/193 ends that failed mapping were rich in repetitive DNA content; with the majority containing interspersed or tandem repeat sequences (183 ends, 95%) and the remaining harbouring segmental duplications (9 ends, 4.7%). ALR content was detected in 7/183

(3.8%) BAC ends that failed mapping due to interspersed or tandem repeat content (Table 7).

Vervet BAC Clone ID	Relative proximity of clone to centromere gap (relative to all fully & half mapped clones)	Mapped end	Chromosome arm	Unmapped end	Repeat content of unmapped end
CH252-030O22	not most proximal	Sp6	2p	T7	ALR
CH252-021L18	most proximal	T7	5p	Sp6	ALR
CH252-030I08	most proximal	Sp6	8q	T7	ALR, GSAT
CH252-049M12	most proximal	T7	8p	Sp6	ALR
CH252-013H22	most proximal	Sp6	11p	T7	ALR, Low complexity:AT rich
CH252-018M20	most proximal	T7	12p	Sp6	ALR
CH252-044I01	most proximal	T7	19q	Sp6	ALR

Table 7: Detection of ALR content in half mapped pericentromeric vervet BAC clones

The majority of these BACs (6/7 cases, 85.6%) also represented vervet clones most proximal (compared to all fully and half mapped clones) to the centromere gaps of human chromosomes (Table 7).

4.5 Insights into old and new centromere organization in the vervet

In Chapter Two, I had identified a total of 15 centromeric locations that appeared to be evolutionary conserved in both the human and the vervet. In human, these locations corresponded to the centromeric regions of chromosomes 5, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19 and the sex chromosomes. To date, I have identified vervet clones (fully mapped, half mapped or both) orthologous to all of the above listed autosomes. Sequence analyses of these clone ends have confirmed that similar to human and other primate species studied, the pericentromeric regions of evolutionary old vervet centromeres are

rich in three major classes of repeats: interspersed and satellite repeats and segmental duplications (Figure 16). Moreover, the presence of alpha-satellite content in the vervet clones mapping most proximally to the centromere gaps suggests that similar to human, the functional domains of evolutionary old vervet centromeres contain alpha-satellite repeats. Identification of a large number (82%, see Chapter Three) of vervet clones believed to be entirely composed of alpha-satellite repeats further supported this hypothesis.

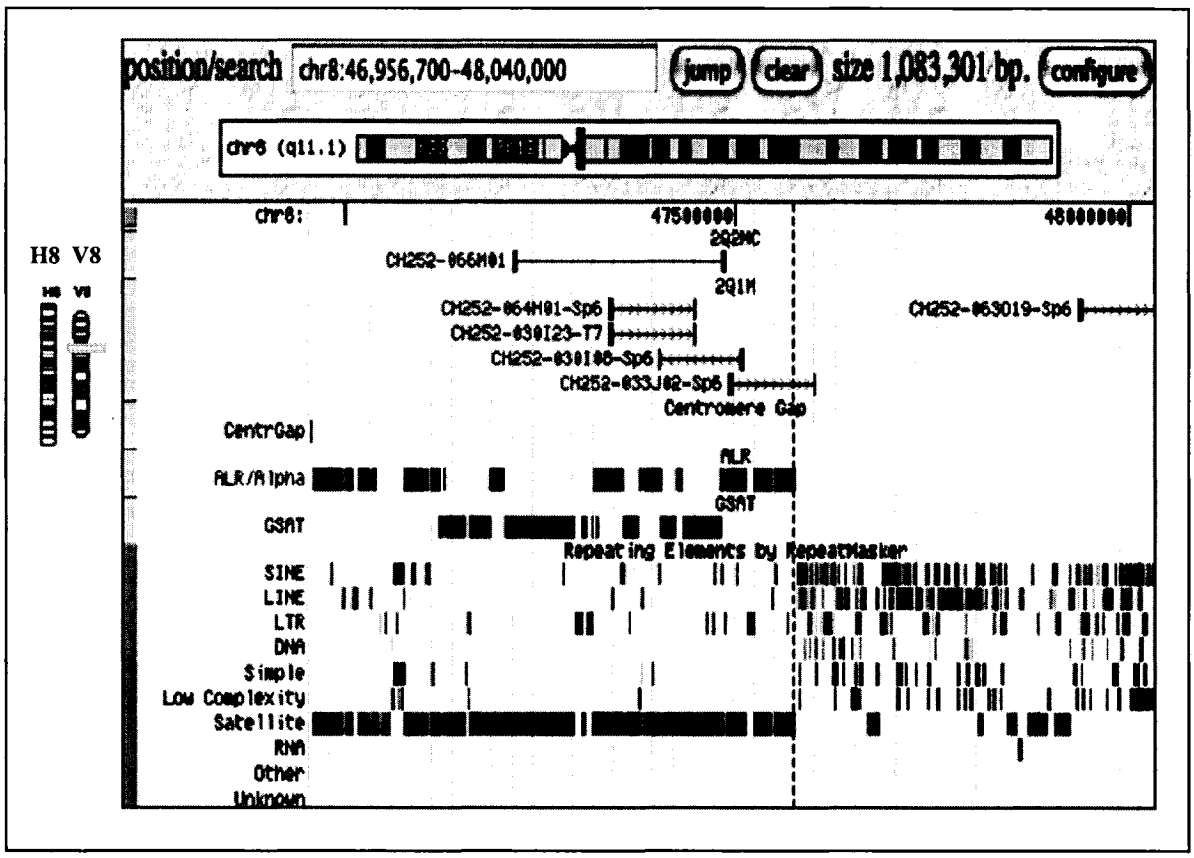

Figure 16: Analyses of an evolutionary old centromere.

The human 8q pericentromeric region displays an abundance of interspersed repeats in the distal pericentromeric region and an enrichment of satellite repeats (ALR and GSAT) in regions proximal to the centromere gaps (Nusbaum et al. 2006). Note the change in repeat classes from one region to the other in human. The perforated black line is used to demarcate the distal and proximal pericentromeric regions of chromosome 8q. End sequence analyses suggest that a similar pattern exists for the vervet 8q pericentromeric region.

However, it should be noted that though the functional domains of evolutionary old centromeres in the vervet is most likely composed of alpha-satellite repeats, the specific alpha-satellite monomers present in these regions might differ from that of the human. My analyses in Chapter Three suggested that the functional domains of vervet centromeres are most likely composed of primordial alpha-satellite monomers, while literature suggests (Alexandrov et al. 2001; Rudd and Willard 2004; Schueler et al. 2005) that in human these regions are populated by more recently evolved alpha-satellite monomer subtypes.

In theory, BAC clones derived from an evolutionary new centromeric region in vervet should appear as a clustering of half-mapped clones at the appropriate location in the human reference map (Figure 17). The unmapped ends in the majority of cases should contain centromeric repeat families, in particular alpha-satellite sequences.

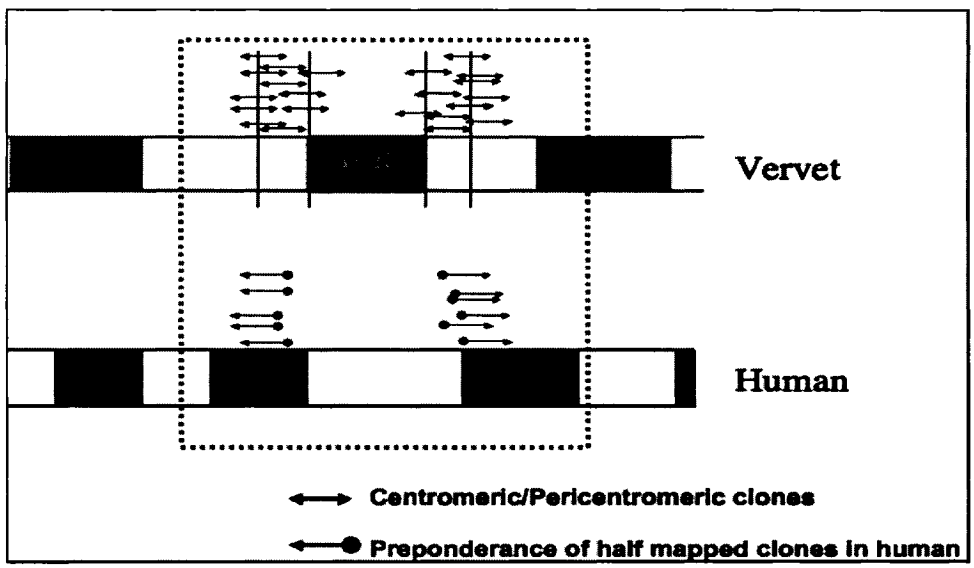

Figure 17: Identifying evolutionary young vervet centromeres using BAC end data

A clustering of vervet BAC clones, at the appropriate location in the human reference map, with one mapped end and the other end unmapped and displaying repeat content, can help map the exact location of new centromere emergence. Blue represents the functional domain of the new centromere structure. The pink lines flanking the functional domain represent the distal pericentromeric boundaries of the new centromere structure. Abbreviations: CEN; centromere.

To date, I have identified five candidate half-mapped clones with alpha-satellite content mapping to a segment of human chromosome 6 (q arm) that roughly corresponds to the location of an evolutionary young centromere in the vervet (vervet chromosome 13 centromere). Though redundant BAC clone coverage is needed to confirm that the identified vervet BACs are actually embedded in a evolutionary new centromeric location, the presence of alpha-satellite content in these BACs is encouraging preliminary evidence that alpha-satellites are present in evolutionary young centromeric locations. It is my belief that evolutionary young centromeres will exhibit a similar organization to that seen in evolutionary old centromeres, though the overall centromeric region might be smaller in size. Literature suggests that the size of centromeric regions increases with age (Schueler et al. 2005). Overall, as the vervet physical map increases in coverage, evolutionary young centromeric regions will become easier to uncover and characterize using BAC end sequence data.

4.6 Summary

In summary, I have demonstrated that vervet BAC clones embedded in pericentromeric regions can be identified by comparative mapping to the human reference map. In total, I identified 501 fully mapped and 193 half mapped vervet BAC clones mapping within the human pericentromeric boundaries. Identification of these vervet pericentromeric/centromeric clones has allowed for the characterization of evolutionary old centromeric regions in the vervet. To date, BAC end sequence data suggests that the

organization of regions flanking evolutionary old centromeric locations in the vervet is similar to that of the human.

Chapter Five

Identification of Rearrangements Using BAC End Data

5.1 Preface

Apart from harbouring evolutionarily recent centromere positions, the vervet genome also harbours other types of genome rearrangements (with respect to the human genome). These rearrangements may affect genetic mapping and physical cloning initiatives and may also be responsible for changes in the vervet's gene organization and genetic make-up. Therefore, in tandem with my analysis of centromeric BACs, I have also developed methods to identify and validate candidate BACs for other classes of rearrangements.

5.2 Inferring regions of genome colinearity

The "electronic mapping" of a clone is predictive of its actual orthologous position when both ends align to a common region with the appropriate distance and orientation between the ends. Our lab has conducted several checks to make sure this holds true for BAC clones mapped by virtue of their unambiguously assigned end sequences. We have deep sequenced electronically mapped BACs from multiple species (cow, bat, rabbit, opossum, platypus and vervet) and confirmed that in all cases the internal sequences match the expected genomic content. In addition, we have screened the vervet BAC library with probes derived from six independent genes. End sequences of

the BACs positive for the probes, in all instances, mapped electronically to locations flanking their corresponding target gene. Based on these observations, it is safe to conclude that with respect to the vervet, concordantly mapping BACs can be used as indicators of genome colinearity.

5.3 Inferring regions of genome rearrangement

Locations of genome reorganization (insertions, deletions, duplications, chromosome breaks, fusions, and transversions) can be discerned employing **(i)** redundant BAC end sequence data and **(ii)** genome rearrangement indicators such as inappropriate clone lengths, parallel sequence orientations and half-mapped or non-mapped BAC clones. This method requires that one of the two genomes being compared be available as a high quality assembly (for example the human genome assembly). In a pilot study using chimp BAC end sequence data and this approach our group was able to confirm the presence of previously known genome rearrangements (a physical map of the chimp and genome already exists) locations between the chimp and human genomes. This method can thus be confidently used to search for synteny breakpoints between the vervet and human genomes.

In Figure 18, I have highlighted the various types of rearrangements that may be inferred from differential alignments of redundant (8X – 10X coverage) vervet BAC ends on the human reference region. The size of rearrangements detected by this method is limited by the average insert size of the cloning vector used.

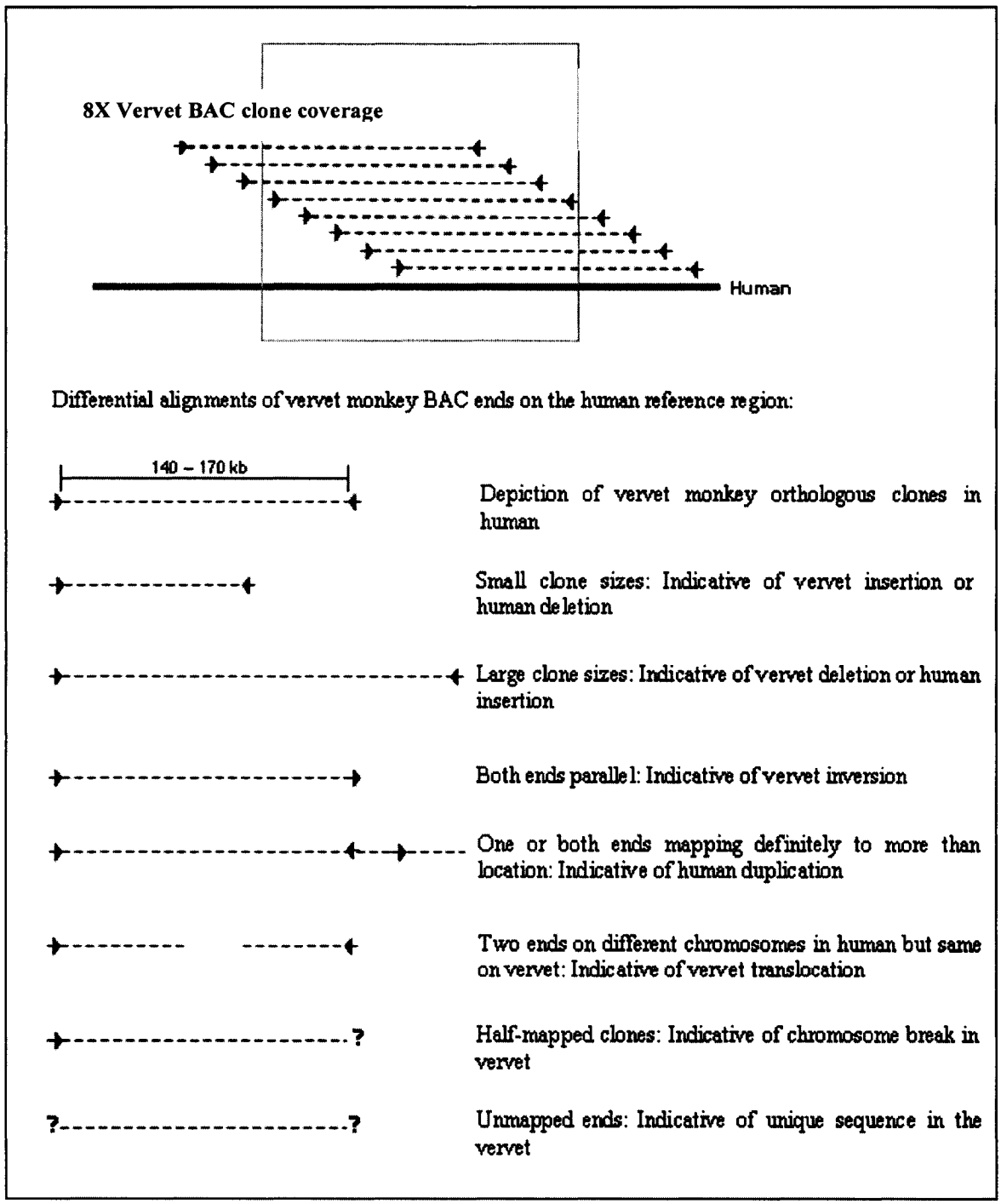

Figure 18: Inferring regions of genome rearrangement

Given a 8X – 10X coverage of the vervet genome, regions of rearrangement between the vervet and the human genomes can be confidently identified using vervet BAC end sequence data. Indicators of genome rearrangement used are inappropriate clone sizes, inappropriate end orientations, discordant mapping results, and the presence of half-mapped and unmapped clones. ' ? ' denotes unmapped BAC end sequences.

For example, the relatively small insert size (~50 kb) of fosmids allow for the detection of rearrangements as small as 15 kb (Tuzun et al. 2005). On the other hand, the substantially larger average insert size of YAC clones considerably reduces the resolution of rearrangement detection. In the case of BAC clones that have an average insert size of approximately 160 kb, only detection of rearrangements bigger than 50 kb in size is predicted.

The current coverage of approximately 0.75X achieved by the vervet physical map project lacks independent data from multiple clones and limits my ability to confidently infer most genome reorganization. Regardless of the low coverage, in sections 5.4 and 5.5 I have identified candidate BAC clones representing two main types of rearrangement events; (1) clones indicating vervet deletion or human insertion and (2) clones indicating translocation events. It should be noted that an individual clone (i.e. a too large clone) might indicate a vervet deletion (or human insertion) event. Inference and confirmation of a vervet translocation event however requires redundant genome coverage.

5.4 Identifying BAC clones indicative of vervet deletion

The BAC cloning procedure employs several rounds of size fractionation resulting in BAC libraries where the range and average insert size of clones can be measured. The CHORI-252 vervet BAC library is reported to have an average clone insert size of ~160 kb, a value determined by pulsed field gel electrophoresis on a sampling of 518 BACs (http://bacpac.chori.org).

5.4.1 Identifying size discordancies

In Figure 19, I have presented the distribution of vervet BAC clone insert sizes determined

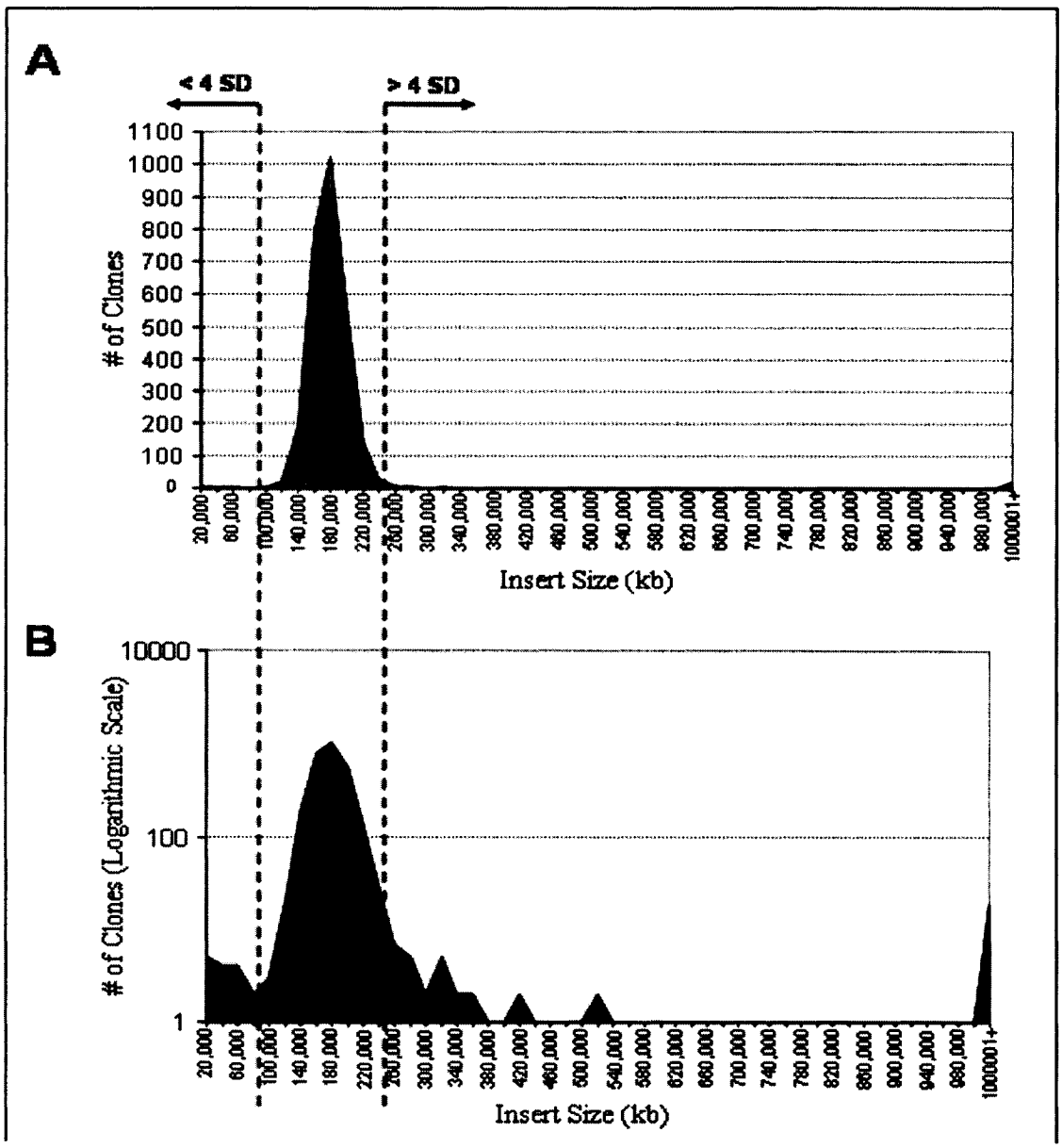

Figure 19: Size distribution of vervet BAC clones

Figure 19A shows the insert size distribution of 2,813 vervet monkey BACs mapping unambiguously to the human genome assembly (build 35). The distance between two end sequences was determined based on the coordinates within the human genome reference. Length thresholds greater or less than four standard deviations (SD) beyond the mean (168 kb) were used to classify length discordancies. Figure 19B shows the same distribution as in Figure 19A, except the clone count is in logarithmic scale thus allowing for better visualization of the number of clones with atypically small and large estimated sizes (± 4 SD from the mean insert size).

from the synteny mapping (on human) coordinates of 2,813 vervet BAC clones. The average insert size calculated from this cohort of electronically mapped vervet BACs was ~168 kb (SD ± 20.7 kb). The slight increase in average insert size (168 kb vs. 160 kb) observed is not surprising and could be attributed to the higher level of interspersed repeat content of the human genome compared to that of the chimpanzee and rhesus, and possibly the vervet. As a consequence, predicted synteny sizes of vervet clones mapping orthologously to the human assembly appear 5% longer than their actual insert size.

I used length thresholds of greater than or less than 4 SD (insert sizes of > 250.6 kb and < 84.8 kb) to classify the predicted insert sizes as atypically small or atypically large. In total, I detected 63 vervet BAC clones that fit this category. To confirm whether these clones were truly atypical in size or represented indicators of genome rearrangements, I selected 18 of these size discordant vervet BACs for laboratory evaluation. 9/18 BAC clones selected had atypically small synteny sizes ranging from 3 kb to 75 kb. The remaining nine clones were atypically large with insert size ranging from 251 kb to 132,447 kb (or 132 Mb) in size.

5.4.2 Validating size discordancies

Purified DNA from the selected BACs containing vector and insert was digested using the enzyme *HindIII*. The resulting fragments were separated by agarose gel electrophoresis (Figure 20, A2 and B2). Gel electrophoresis is a common laboratory technique used to separate DNA fragments according to size by applying an electric current to them. The overall size of each BAC clone was determined by taking into account the number of bands detected and the estimated size of each band.

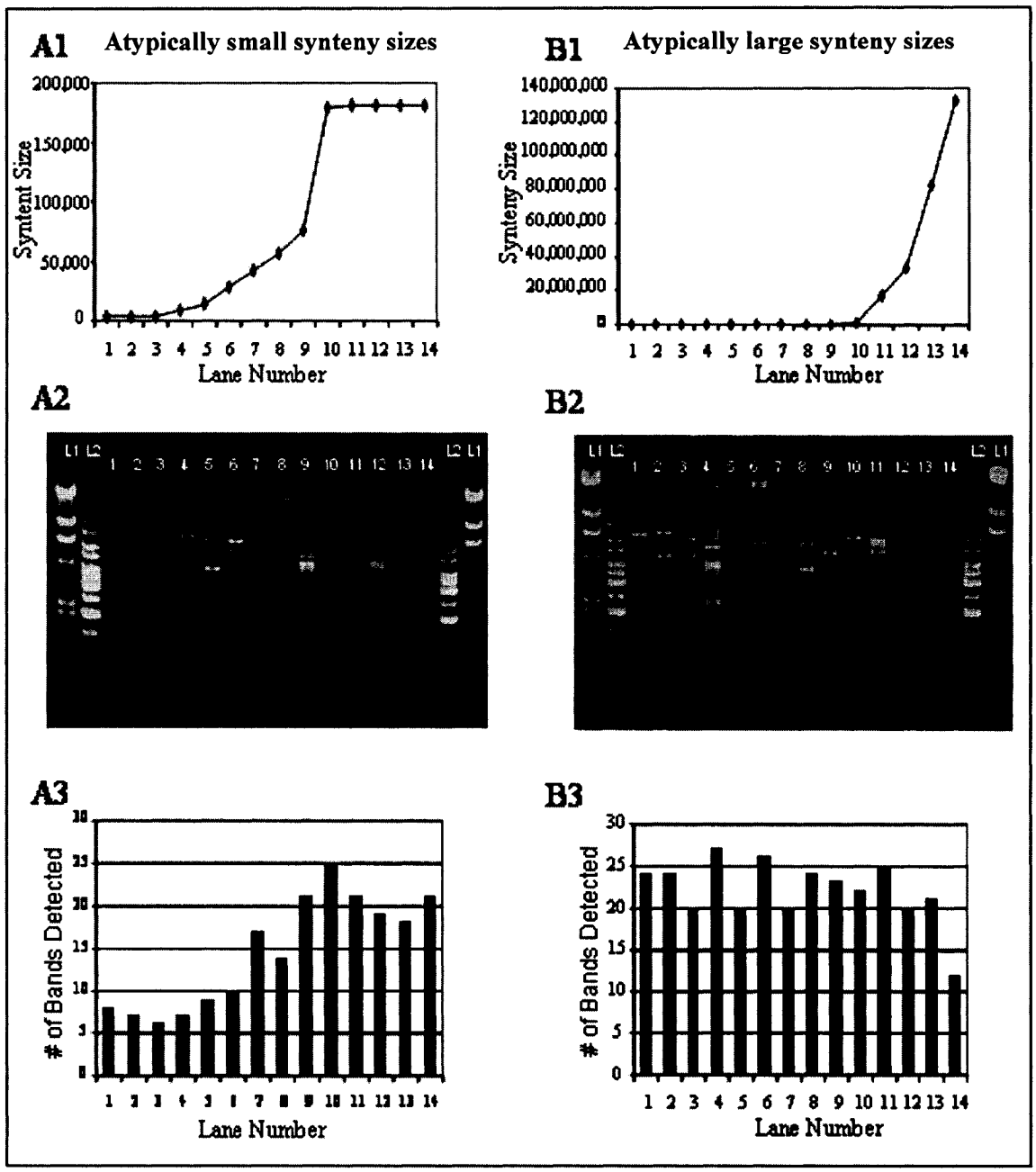

Figure 20: Verifying estimated clone size using agarose gel electrophoresis.

Graphs A1 and B1 show the clone sizes predicted by synteny mapping. Photos A2 and B2 display the agarose gel electrophoresis bands of the *HindII* digested clones. Graphs A3 and B3 displays the number of bands observed in lanes 1-14 of the agarose gels shown in A2 and B2.

Results indicate that 8/9 BACs displaying small synteny sizes in actuality did harbour small clones as evident by the detection of fewer bands compared to the controls

(Figure 20, A3 and B3). These findings led me to conclude that an individual BAC with a predicted small insert size is not a reliable indicator of vervet genome rearrangements (vervet insertions/human deletions) by itself. On the other hand, no major difference in band number was detected between the BACs with atypically large insert sizes and controls. This suggests that individual vervet BACs exhibiting large (>4 SD) predicted insert sizes and mapping unambiguously to the human reference could reliably point to regions of genome evolution (vervet deletion/human insertion).

5.5 Identifying vervet BAC clones indicative of translocation events

To date, I have detected 172 vervet BACs where each end maps unambiguously to separate chromosomes on human. Because the forward and reverse sequence reads from each BAC are physically linked in the vervet genome, placement of each of the two ends to a separate human chromosome might point to a translocation event in either the human or vervet lineages. A translocation event is defined as the "breakage and removal of a large segment of DNA from one chromosome, followed by the segment's attachment to a different chromosome" (www.genome.gov/glossary.cfm). Translocations can be reciprocal, non-reciprocal, or Robertsonian.

5.5.1 Clustering of vervet inter-chromosomal clones

The distribution shown in Figure 21 reveals that chromosomes 1, 4, 6, 10 and X contain a high number (> 25) of mapped ends belonging to independent vervet inter-chromosomal BAC clones. It should be noted that human chromosomes 1, 4 and 6 have all undergone fission in the vervet lineage (discussed in Chapter Two).

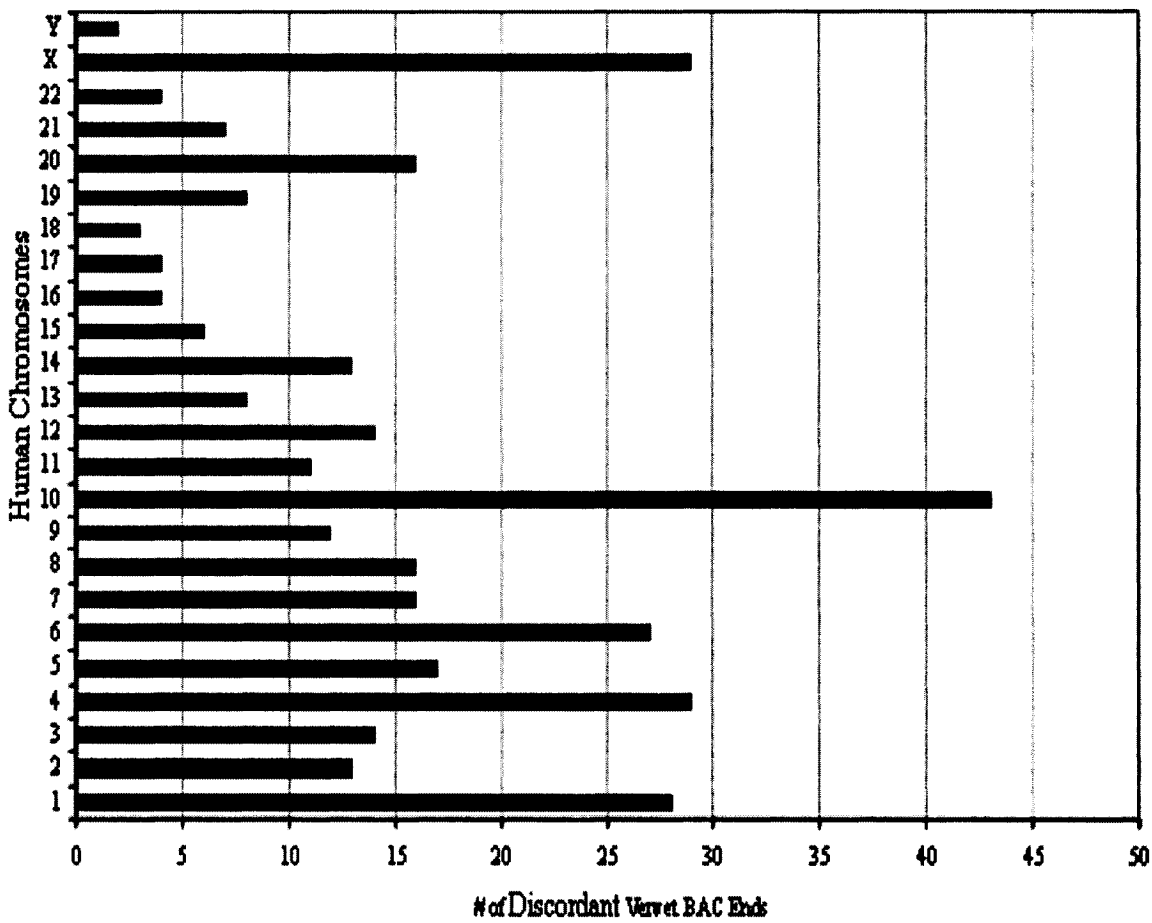

Figure 21: Distribution of inter-chromosomal vervet clone ends across the human genome

This graph summarizes per human chromosome the frequency of vervet BAC clones where the other end maps to a different human chromosome.

Intriguingly, 4/6 (67%) inter-chromosomal vervet BACs mapping to human chromosomes 6 and 9 displayed mapping to pericentromeric regions of chromosome 6, while their corresponding ends mapped to the telomeric region (p arm) of human chromosome 9. Similarly, the majority of such ends aligning to the telomeric regions (p arm) of human chromosome 10 had their corresponding ends mapping to either human chromosome 1 or 4.

These observations led me to hypothesize that in the vervet, fission chromosomes might have acquired telomeres by capturing terminal fragments from other chromosomes via mechanisms of non-reciprocal translocation (Bosco and Haber 1998; Shaffer and Lupski 2000; Maser and DePinho 2004). In human, such translocation events have been known to mediate telomere acquisition in tumour cells (Meltzer et al. 1993; Lo et al. 2002). Telomere capping is extremely important in eukaryotes, since they protect chromosomes from enzymatic end-degradation and maintain stability. "Chromosomes with truncated telomeric tips fuse with other chromosome ends or become lost during cell division" (Blackburn 2005). Telomeric capping is thus essential for chromosomes undergoing fission to survive as two independent chromosomes.

5.5.2 Rapid identification of clustering inter-chromosomal clones

Presented in Figure 22 is the output from an automated visualization tool that allows for the rapid identification of clustering inter-chromosomal clones. Discordant BACs are tagged as clustering if the ends of two or more independent clones map within a 1 Mb region of one chromosome and the corresponding ends map within a 1 Mb region of another chromosome. Already of use at our current coverage of 0.75X, this visualization tool will help discern locations of translocation events as the vervet physical map increases in coverage.

Taken together my findings suggest that in the vervet genome one can begin to see clustering of discordant BAC clones even at low coverages. This feature can probably be attributed to the uniqueness of the vervet genome when compared to human and most other primates, in particular with respect to chromosome fission events.

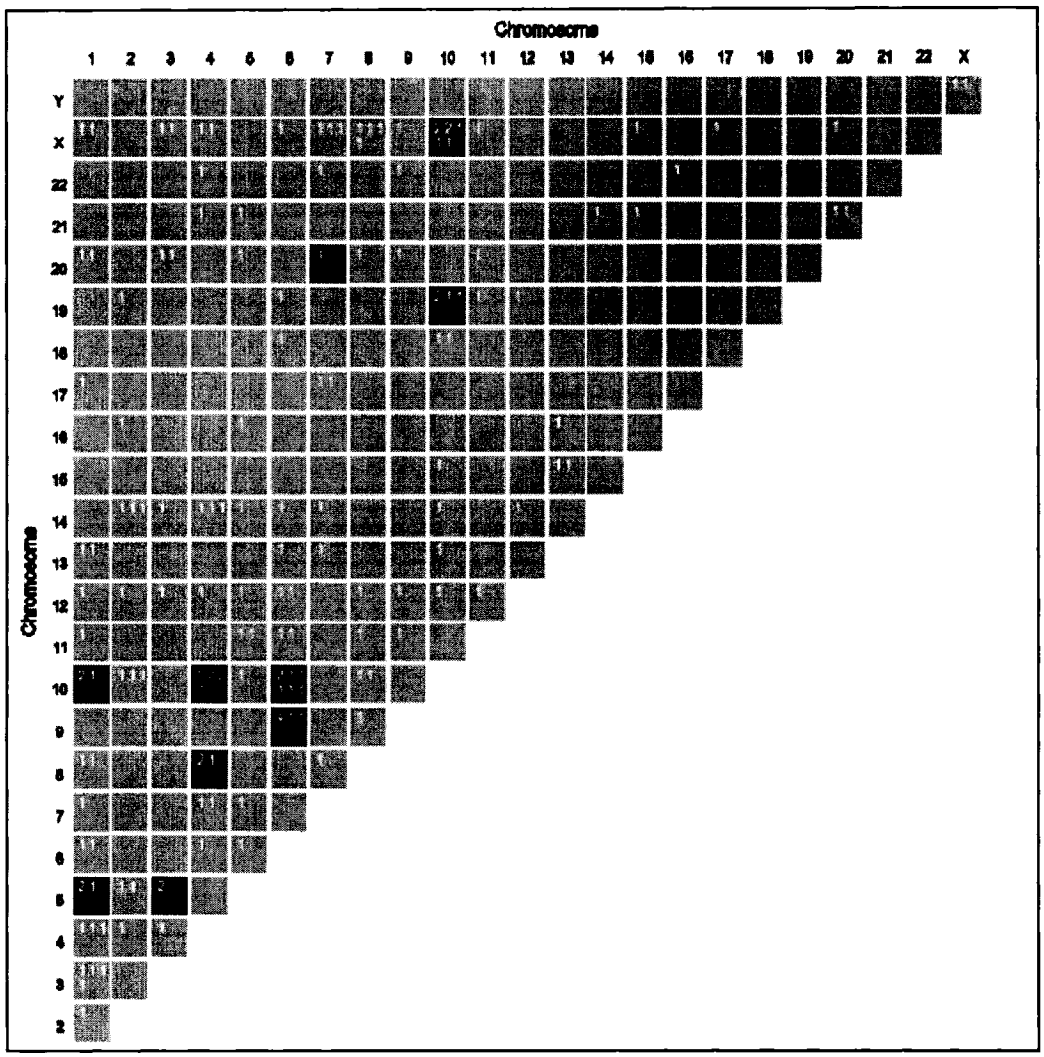

Figure 22: Clustering of discordant clones between chromosomes

Colour of the squares range from light green to bright red. A light green colour represents an absence of clustering discordant clones. Dark green, brown, dark red and bright red squares indicate the clustering of 2, 3, 4 and 5 independent discordant clones respectively, within a 1 Mb region in each of the two chromosomes involved. The sum of the numbers within each square provides the total number of discordant clones between the two chromosomes involved. For example, the total number of discordant clones between chromosome 4 and chromosome 10 is calculated as 5 + 2 + 2 + 1 = 10.

5.6 Summary

Overall, I have established that the average synteny size of concordantly mapped vervet BAC clones is approximately 168 kb. Vervet BAC clones demonstrating synteny sizes greater than 250 kb appear to be good indicators of vervet deletion or human

insertion events, even at low coverages. In addition, I have demonstrated that clustering of unambiguously mapping inter-chromosomal vervet BAC ends can been seen even at a low map coverage of 0.75X. This finding suggests that compared to the human genome, the vervet genome harbours a greater number of translocation events, particularly at chromosome fission breakpoints. I have also developed a method for rapid identification of clustering inter-chromosomal clones.

Chapter Six

Discussion and future research opportunities

Several mechanisms can lead to the reorganization of genomes during speciation. I have demonstrated using a human/rhesus/vervet karyotype comparison that among the primates, the vervet genome contains a remarkable number of fission chromosomes. Fission chromosomes may arise as a result of centromere repositioning, new centromere emergence or other rearrangement events. My results using comparative karyotype analyses suggest that the vervet genome contains at least 12 evolutionary new centromere locations. Based on the estimated divergence time between the vervet and the rhesus (Goodman 1999), these new centromeric locations are inferred to have emerged less than 9 million years ago.

The evidence that there are numerous evolutionary new centromere locations in the vervet provides a unique opportunity to study the evolutionary dynamics of centromere formation. To pursue this, I developed two methodologies to identify vervet BAC inserts belonging to centromeric regions. In the first approach, I analysed the human and chimp genomes to identify a marker that was specific to and present at all centromeres, and determined that the alpha-satellite repeat satisfied these conditions. I confirmed that this marker was present in the vervet genome using both sequence analysis and hybridization screening. Vervet BAC clones were classified as deep centromeric if both end sequences solely contained alpha-satellite monomers, as this was indicative of similar repeat content throughout. Vervet BAC inserts that lacked exclusive alpha-

satellite content on both ends were binned as pericentromeric or as belonging to short alpha-satellite islands.

A large body of work associates long arrays of alpha-satellite repeats with primate centromere function (Schueler et al. 2001; Spence et al. 2002; Grimes et al. 2004; Rudd et al. 2006). The functional domains of human centromeres are composed of highly ordered arrays of evolutionary young alpha-satellite monomers (Alexandrov et al. 2001; Schueler et al. 2001). To date, the existence of such monomers had not been reported in the vervet. A novel finding of my thesis research project is the detection of these evolutionary young alpha-satellite monomers in the vervet genome. The majority of vervet BACs containing evolutionary recent alpha-satellite monomers were classified as pericentromeric since their end sequences did not exclusively contain alpha-satellite repeats. This finding points to the possibility that the alpha-satellite monomer organization in the vervet is different from that of the human, with evolutionary young monomers in the vervet primarily residing in pericentromeric locales.

In my second approach to identifying vervet BAC clones belonging to centromeric regions, I demarcated the pericentromeric boundaries in human and identified vervet BAC inserts that mapped orthologously to these regions. Mapping of vervet clones to centromeric locations predicted to be evolutionary conserved in both species lend support to the hypothesis that these centromeres were present in the primate ancestor that gave rise to the ape and the Old World monkey lineages. Since the ape and Old World monkey lineages diverged ~25 million years ago (Goodman 1999), evolutionary old centromeres in the vervet appear to be greater than 25 million years old.

In addition, vervet BAC end data suggest that the sequence content and organization at evolutionary old pericentromeric regions have remained largely unchanged in the two species. Not unlike other primate species studied (Schueler et al. 2005), evolutionary old pericentromeric region in the vervet is fertile in repeat content belonging to three major classes; interspersed, satellite and segmental duplications. This is not surprising since the accumulation of repetitive DNA at centromeric locations "allows such sites in the genome to function more competently as centromeres" (Eichler 1999). The functional competency and conservation of centromeric locations in primate genomes is dependent on the ability of these DNA segments to replicate last during meiosis. Repetitive DNA has a stalling effect on the DNA replication fork that in effect prolongs replication time and ensures the replication of centromeric regions in late meiosis (Samadashwily et al. 1997; Eichler 1999; Regelson et al. 2006).

Since accumulation of repetitive DNA in regions flanking the functional centromere is a temporal phenomenon (Barry et al. 1999; Eichler 1999; Henikoff 2002), I anticipate that new centromeric locations in the vervet will exhibit a shorter stretch of pericentromeric DNA. To date, I have identified a few candidate BAC clones from sites of new centromere emergence in the vervet. However characterization of the long-range organization of evolutionary young centromeric regions in the vervet requires the identification of additional BACs from such regions.

In addition to harbouring evolutionarily recent centromeres, the vervet genome also houses other types of genome rearrangements. I developed methodologies to detect candidate centromeric and whole genome rearrangements, in particular vervet deletion/human insertion and vervet translocation events. I demonstrated that a single

vervet BAC clone can easily detect vervet deletion/human insertion events greater can 80 kb in size (BAC sizes > 4 SD). As the BAC clone coverage of the vervet genome increases, it will become possible to detect smaller vervet deletion/human insertion event. Using reiterative simulated data I calculated that a 6X coverage of the vervet genome will allow for the detection of vervet deletion/human insertion events as small as 15 kb in size.

Overall, the various tools and approaches that I have developed to study genome evolution will prove useful as the vervet monkey project ramps up. The many insights gained from this project regarding vervet genome evolution will allow for a more refined mapping of vervet chromosome correspondences and lead to the identification of additional rearranged genomic locales.

6.1 Future research opportunities

Vervet BACs can be localized to chromosomal regions using fluorescent in situ hybridization (FISH) experiments. FISH is a cytogenetic technique that employs fluorescent probes to detect and localize DNA sequences on chromosomes. To date, I have electronically mapped several vervet BAC clones to orthologous pericentromeric regions in human. I would propose that FISH be performed on selected BACs from this group to confirm localization of the clones to vervet pericentromeric regions. This technique provides us a means to study change in sequence content (ALR and other repeats), as one approaches the functional centromere. The sequenced clone could then be a source of probes (or sequences to align with BAC ends) to progressively walk in towards the centromere.

In addition, I have identified multiple vervet BACs where both ends solely contain alpha-satellite sequences. These BAC inserts are assumed to be entirely composed of alpha-satellite sequences and belonging to deep centromeric regions. Survey sequencing or deep sequencing of selected vervet BAC clones should prove or disprove this assumption, and potentially provide a more refined placement for these vervet BAC clones.

Appendix 1:

Perl Script

IntersectBacPairs.pl

```
#!/usr/bin/perl –w

use strict;
my $bacendfile = $ARGV[0];
my $regions = $ARGV[1];
my $rptotal = 0;
my $cttotal = 0;

print "$bacendfile\n";
print "$regions\n";


open (LOC, $regions);

while (my $line=<LOC>){
        print $line;
        my @data;
        @data = split("\t", $line);
        my $chr;
        my $location;
        $chr = $data[0];
        $location = $data[1];
        chomp $location;
        print "$chr, $location\n";
        open (PB, $bacendfile);
        my $rpcount = 0;
        my $ctcount = 0;
        while (my $bacline=<PB>) {
                #print $bacline;
                my @bacdata;
                @bacdata = split("\t", $bacline);
                my $bacchr;
                my $bacstart;
                my $bacend;
                my $bacname;
                $bacchr = $bacdata[1];
```

```
                $bacstart = $bacdata[2];
                $bacend = $bacdata[3];
                $bacname = $bacdata[4];
                $bacchr =~ s/chr//;
                        if (($chr eq $bacchr) and
                        ($bacstart <= $location) and
                        ($bacend >= $location)) {
                        print "$bacchr, $bacstart, $bacend, $bacname ";
                                if ($bacname =~/RP.*/){
                                 $rpcount = $rpcount + 1;
                                 $rptotal = $rptotal +1;
                                }
                                if ($bacname =~/CT.*/){
                                 $ctcount = $ctcount + 1;
                                 $cttotal =$cttotal + 1;
                                }
                        print "\n";
                        }

        }

        close (PB);
print "RP = $rpcount, CT = $ctcount\n";
}

close (LOC);
print "RP_Total = $rptotal, CT_Total = $cttotal\n";

exit;
```

Hybridization of Overgo Probes to BAC filters

This original protocol was developed by Dr. John D. McPherson (Washington University School of Medicine, Genome Sequencing Center).

Reagents

Hybridization solution

1 mM EDTA

7% SDS

0.5 M sodium phosphate (pH 7.2)

Church buffer

Hybridization solution plus 1% BSA (fraction V)

1M sodium phosphate, pH 7.2

426 g sodium phosphate dibasic, anhydrous (Na_2HPO_4) in 1700 ml H_2O, add 12 ml 85% H_3PO_4 and make to 3 L

0.5 M EDTA, pH 8.0

200 g EDTA - 4Na and 176.44 g EDTA - 2Na in 1,600 ml H_2O, pH to 8.0 with NaOH pellet and make to 2 L

Washing buffer B

1 mM EDTA, 1 % SDS, 40 mM Na_2HPO_4

12 ml 0.5 M EDTA

60 g SDS

240 ml 1 M Na_2HPO_4, pH 7.2

to 6 L with ddH_2O

Washing solution 2

1.5x SSC, 0.1% SDS

375 ml 20x SSC

25 ml 20% SDS

to 5 L with ddH_2O

Washing solution 3

0.5x SSC, 0.1% SDS

125 ml 20x SSC

25 ml 20% SDS

to 5 L with ddH_2O

STE buffer

10 mM Tris-HCl, pH 8.0

1 mM EDTA

0.438 g NaCl

total volume is 50 ml

20x SSC

701.2 g NaCl

352.8 g sodium citrate, dihydrate

pH to 7.0 with a few drop of a conc. HCl

Bring volume to 4 L with ddH2O

20% SDS

400g SDS (Ultra Pure: Life Technologies) / 2 L ddH2O

BSA (2 mg/ml)

Dilute 10ml/ml BSA (purchased from New England Biolabs) with ddH2O prior to use

Solution O

1.25 M Tris-HCl, pH 8.0, 125 mM MgCl2

625 µl 2 M Tris-HCl (pH8.0)

125 ul 1 M MgCl2

250 ul ddH2O

Solution A

1 ml solution O

18 ul 2-mercaptoethanol

5 ul 100 mM dTTP

5 ul 100 mM dGTP

Solution B

2 M HEPES-NaOH, pH 6.6

Dissolve 23.8g of HEPES-free acid (molecular weight 238.3; SIGMA H-4034) in 40 ml de-ionized distilled water and adjust the pH to 6.6 with NaOH. Adjust the final volume to 50 ml with deionized distilled water.

Solution C

3 mM Tris-HCl, pH 7.4, 0.2 mM EDTA

15 ul 1 M Tris-HCl (pH7.4)

2 ul 0.5 M EDTA

to 5 ml with ddH2O

Prepare OLB: A:B:C, 1:2.5:1.5

Solution A 1 ml

Solution B 2.5 ml

Solution C 1.5 ml

Aliquot and store at -20 C.

Overgo Labeling

0.5 ul of complementary 20 uM oligos (0.25 ul of each) was combined with 4.5 ul ddH2O and heated to 80° C and 37° C for 5 and 10 minutes respectively. The oligo mix was then rapidly cooled on ice and flash spinned. 0.5 ul BSA (2 mg/ml), 2.0 ul OLB (-A, -C), 0.5 ul 32P-dATP, 0.5 ul 32P-dCTP, and 1 ul Klenow fragment (2 U/ul) was added to the oligo mix. The resulting labeling solution was incubated at room temperature for 2 hours and diluted to 50 ul using STE buffer. Labelled probes were purified using ProbeQuant G-50 micro column. The probes were subsequently heated at 95° C for 10 minutes; cooled on ice, briefly centrifuged, mixed and divided into 20 ul aliquots.

Hybridization of Overgos to BAC filters

Filters were pre-hybridized for 1 hour with 50 ml of warmed hybridization solution (58° C to 65° C). Labeled probes were then added and allowed to hybridize with filters overnight. After hybridization, membranes were washed in solutions with increasing stringency starting with cold 2× SSC, followed by warm 2× SSC, warm 1.5× SSC and 0.5× SSC respectively. A final rinse was performed using cold 2× SSC. Membranes were then sealed in plastic wrap and exposed to film with an intensifying screen over the film at −80° C for 1 day.

Bibliography

References

Alexandrov I, Kazakov A, Tumeneva I, Shepelev V, Yurov Y (2001) Alpha-satellite DNA of primates: old and new families. Chromosoma 110:253-266.

Amor DJ, Choo KH (2002) Neocentromeres: role in human disease, evolution, and centromere study. Am J Hum Genet 71:695-714.

Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X (2003) Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. Hum Mol Genet 12:2201-2208.

Bailey J, Jorgensen M, Fairbanks L (2001a) Introducing the UCLA-VA Vervet Monkey Research Colony as a genetic resource. Paper presented at Human Genome Meeting. Edinburg. http://hgm2001.hgu.mrc.ac.uk/Abstracts/Publish/WorkshopsPoster/WorkshopPoster07/hgm0208.htm

Bailey JA, Church DM, Ventura M, Rocchi M, Eichler EE (2004) Analysis of segmental duplications and genome assembly in the mouse. Genome Res 14:789-801.

Bailey JA, Liu G, Eichler EE (2003) An Alu transposition model for the origin and expansion of human segmental duplications. Am J Hum Genet 73:823-834.

Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE (2001b) Segmental duplications: organization and impact within the current human genome project assembly. Genome Res 11:1005-1017.

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M, Eichler EE (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. Am J Hum Genet 70:83-100

Barry AE, Howman EV, Cancilla MR, Saffery R, Choo KH (1999) Sequence analysis of an 80 kb human neocentromere. Hum Mol Genet 8:217-227.

Batzoglou S, Berger B, Mesirov J, Lander ES (1999) Sequencing a genome by walking with clone-end sequences: a mathematical analysis. Genome Res 9:1163-1174.

Baxevanis AD, Ouellette BFF (2005) Bioinformatics: A practical Guide to the analysis of genes and proteins. John Wiley & Sons, Inc.

Bellino FL, Wise PM (2003) Nonhuman primate models of menopause workshop. Biol Reprod 68:10-18.

Bier E, McGinnis W (2003) Model Organisms in the Study of Development and Disease. In: Epstein CJ, Erickson RP, Wynshaw-Boris A (eds) Inborn Errors of Development: The Molecular Basis of Clinical Disorders of Morphogenesis. Oxford University Press, Oxford, pp 25-45.

Blackburn EH (2005) Telomeres and telomerase: their mechanisms of action and the effects of altering their functions. FEBS Lett 579:859-862.

Bosco G, Haber JE (1998) Chromosome break-induced DNA replication leads to nonreciprocal translocations and telomere capture. Genetics 150:1037-1047.

Boulton AM, Horrocks JA, J. B (1996) The Barbados vervet (Cercopithecus aethiops sabaeus): changes in population size and crop damage, 1980-1994. Int J Primatol 17(5):831-844.

Carlsson HE, Schapiro SJ, Farah I, Hau J (2004) Use of primates in research: a global overview. Am J Primatol 63:225-237.

Cawthon Lang KA (2006) Primate Factsheets: Vervet (Chlorocebus) Taxonomy, Morphology, & Ecology. http://pin.primate.wisc.edu/factsheets/entry/vervet. Accessed August 3, 2006 .

Chan AW (2004) Transgenic nonhuman primates for neurodegenerative diseases. Reprod Biol Endocrinol 2:39.

Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive DNA in eukaryotes. Nature 371:215-220.

Cheng Z, Ventura M, She X, Khaitovich P, Graves T, Osoegawa K, Church D, DeJong P, Wilson RK, Paabo S, Rocchi M, Eichler EE (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. Nature 437:88-93.

Choo KH (1997) Centromere DNA dynamics: latent centromeres and neocentromere formation. Am J Hum Genet 61:1225-1233.

De Grouchy J (1987) Chromosome phylogenies of man, great apes, and Old World monkeys. Genetica 73:37-52.

Deaner RO, Khera AV, Platt ML (2005) Monkeys pay per view: adaptive valuation of social images by rhesus macaques. Curr Biol 15:543-548.

Dubchak I, Frazer K (2003) Multi-species sequence comparison: the next frontier in genome annotation. Genome Biol 4:122.

Eichler EE (1999) Repetitive conundrums of centromere structure and function. Hum Mol Genet 8:151-155.

Eichler EE (2001) Segmental duplications: what's missing, misassigned, and misassembled--and should we care? Genome Res 11:653-656.

Eichler EE, DeJong PJ (2002) Biomedical applications and studies of molecular evolution: a proposal for a primate genomic library resource. Genome Res 12:673-678.

Eichler EE, Frazer KA (2004) The nature, pattern and function of human sequence variation. Genome Biol 5:318.

Emanuel BS, Shaikh TH (2001) Segmental duplications: an 'expanding' role in genomic instability and disease. Nat Rev Genet 2:791-800.

Enard W, Paabo S (2004) Comparative primate genomics. Annu Rev Genomics Hum Genet 5:351-378.

Ervin F, Palmour R (2003) Primates for 21st Century Biomedicine: The St. Kitts Vervet (Chlorocebus aethiops, SK). In: International Perspectives: The Future of Nonhuman Primate Resources, Proceedings of the Workshop Held April 17-19, 2002. Natl Acad Pr., pp 49-53.

Ervin FR, Palmour RM, Young SN, Guzman-Flores C, Juarez J (1990) Voluntary consumption of beverage alcohol by vervet monkeys: population screening, descriptive behavior and biochemical measures. Pharmacol Biochem Behav 36:367-373.

Fairbanks L (2003) Demands for Rhesus Monkeys in Biomedical Research: A Workshop Report: Alternative Old World Primate Models for Non-AIDS Research African Green Monkeys (Vervets). ILAR 44:230-231.

Fedigan L, Fedigan LM (1988) Cercopithecus aethiops: a review of field studies. . In: Gautier-Hion A, Bourlière F, Gautier JP, Kingdon J (eds) A primate radiation: evolutionary biology of the African guenons Cambridge Univ Pr., Cambridge (UK), pp 387-411.

Finelli P, Stanyon R, Plesker R, Ferguson-Smith MA, O'Brien PC, Wienberg J (1999) Reciprocal chromosome painting shows that the great difference in diploid number between human and African green monkey is mostly due to non-Robertsonian fissions. Mamm Genome 10:713-718.

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC (2003) Cross-species sequence comparisons: a review of methods and available resources. Genome Res 13:1-12.

Garcia-Castells E, Juarez Gonzalez J, Ervin FR, Guzman-Flores C (1989) Changes in social dynamics associated to the menstrual cycle in the vervet monkey (Cercopithecus aethiops). Bol Estud Med Biol 37:11-16.

Goodman M (1999) The genomic record of Humankind's evolutionary roots. Am J Hum Genet 64:31-39.

Grimes BR, Babcock J, Rudd MK, Chadwick B, Willard HF (2004) Assembly and characterization of heterochromatin and euchromatin on human artificial chromosomes. Genome Biol 5:R89.

Guy J, Hearn T, Crosier M, Mudge J, Viggiano L, Koczan D, Thiesen HJ, Bailey JA, Horvath JE, Eichler EE, Earthrowl ME, Deloukas P, French L, Rogers J, Bentley D, Jackson MS (2003) Genomic sequence and transcriptional profile of the boundary between pericentromeric satellites and genes on human chromosome arm 10p. Genome Res 13:159-172.

Hall AE, Keith KC, Hall SE, Copenhaver GP, Preuss D (2004) The rapidly evolving field of plant centromeres. Curr Opin Plant Biol 7:108-114.

Hau J, Farah IO, Carlsson H, Hagelin J (2000) Opponents' Statement: non-human primates must remain accessible for vital biomedical research. In: Balls M, van Zeller AM, Halder M (eds) Progress in the Reduction, Refinement and Replacement of Animal Experimentation, Developments in Animal and Veterinary Sciences. Vol 31B. Elsevier, Oxford, pp 1593-1601.

Hau J, Schapiro SJ (2006) Non-human Primates in Biomedical Research. Scand J Lab Anim Sci 33:9 - 12.

Henikoff S (2002) Near the edge of a chromosome's "black hole". Trends Genet 18:165-167.

Henikoff S, Dalal Y (2005) Centromeric chromatin: what makes it unique? Curr Opin Genet Dev 15:177-184.

Henikoff S, Malik HS (2002) Centromeres: selfish drivers. Nature 417:227.

Horvath JE, Schwartz S, Eichler EE (2000a) The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. Genome Res 10:839-852.

Horvath JE, Viggiano L, Loftus BJ, Adams MD, Archidiacono N, Rocchi M, Eichler EE (2000b) Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. Hum Mol Genet 9:113-123.

Jauch A, Wienberg J, Stanyon R, Arnold N, Tofanelli S, Ishida T, Cremer T (1992) Reconstruction of genomic rearrangements in great apes and gibbons by chromosome painting. Proc Natl Acad Sci U S A 89:8611-8615.

Jones S, Martin R, Pilbeam D (1992) The Cambridge Encyclopedia of Human Evolution. Cambridge University Press.

Kaessmann H, Wiebe V, Weiss G, Paabo S (2001) Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat Genet 27:155-156.

Kim UJ, Shizuya H, de Jong PJ, Birren B, Simon MI (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res 20:1083-1085.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921.

Lemere CA, Beierschmitt A, Iglesias M, Spooner ET, Bloom JK, Leverone JF, Zheng JB, Seabrook TJ, Louard D, Li D, Selkoe DJ, Palmour RM, Ervin FR (2004) Alzheimer's

disease abeta vaccine reduces central nervous system abeta levels in a non-human primate, the Caribbean vervet. Am J Pathol 165:283-297.

Liberini P, Pioro EP, Maysinger D, Ervin FR, Cuello AC (1993) Long-term protective effects of human recombinant nerve growth factor and monosialoganglioside GM1 treatment on primate nucleus basalis cholinergic neurons after neocortical infarction. Neuroscience 53:625-637.

Lo AW, Sprung CN, Fouladi B, Pedram M, Sabatier L, Ricoul M, Reynolds GE, Murnane JP (2002) Chromosome instability as a result of double-strand breaks near telomeres in mouse embryonic stem cells. Mol Cell Biol 22:4836-4850.

Locke DP, Segraves R, Carbone L, Archidiacono N, Albertson DG, Pinkel D, Eichler EE (2003) Large-scale variation among human and great ape genomes determined by array comparative genomic hybridization. Genome Res 13:347-357.

Maggert KA, Karpen GH (2001) The activation of a neocentromere in Drosophila requires proximity to an endogenous centromere. Genetics 158:1615-1628.

Martin S, Palmour RM, Goldwater R, Gutkowsa J, Hughes C, Hamet P, Ervin FR (1990) Characterization of a primate model of hypertension. The response of hypertensive and normotensive male vervets (Cercopithecus aethiops) to cold pressor stress, captopril administration, and acute bolus of atrial natriuretic factor. Am J Hypertens 3:27-32.

Maser RS, DePinho RA (2004) Telomeres and the DNA damage response: why the fox is guarding the henhouse. DNA Repair (Amst) 3:979-988.

Meltzer PS, Guan XY, Trent JM (1993) Telomere capture stabilizes chromosome breakage. Nat Genet 4:252-255.

Mikkelsen TS, Hillier LW, Eichler EE, Zody MC, Jaffe DB, Yang S, Enard W, et al. (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437:69-87.

Milosavljevic A, Harris RA, Sodergren EJ, Jackson AR, Kalafus KJ, Hodgson A, Cree A, Dai W, Csuros M, Zhu B, de Jong PJ, Weinstock GM, Gibbs RA (2005) Pooled genomic indexing of rhesus macaque. Genome Res 15:292-301.

Montefalcone G, Tempesta S, Rocchi M, Archidiacono N (1999) Centromere repositioning. Genome Res 9:1184-1188.

Mrasek K, Heller A, Rubtsov N, Trifonov V, Starke H, Claussen U, Liehr T (2003) Detailed Hylobates lar karyotype defined by 25-color FISH and multicolor banding. Int J Mol Med 12:139-146.

Muller S, O'Brien PC, Ferguson-Smith MA, Wienberg J (1997) Reciprocal chromosome painting between human and prosimians (Eulemur macaco macaco and E. fulvus mayottensis). Cytogenet Cell Genet 78:260-271.

Muller S, Stanyon R, O'Brien PC, Ferguson-Smith MA, Plesker R, Wienberg J (1999) Defining the ancestral karyotype of all primates by multidirectional chromosome painting between tree shrews, lemurs and humans. . Chromosoma 108:393-400.

Muller S, Wienberg J (2001) "Bar-coding" primate chromosomes: molecular cytogenetic screening for the ancestral hominoid karyotype. Hum Genet 109:85-94.

Murphy WJ, Agarwala R, Schaffer AA, Stephens R, Smith C, Jr., Crumpler NJ, David VA, O'Brien SJ (2005) A rhesus macaque radiation hybrid map and comparative analysis with the human genome. Genomics 86:383-395.

Nadon NL (2006) Of mice and monkeys: national institute on aging resources supporting the use of animal models in biogerontology research. J Gerontol A Biol Sci Med Sci 61:813-815.

Nobrega MA, Pennacchio LA (2003) Comparative genomic analysis as a tool for biological discovery. J Physiol 554.1:31-39.

Nusbaum C, Mikkelsen TS, Zody MC, Asakawa S, Taudien S, Garber M, Kodira CD, et al. (2006) DNA sequence and analysis of human chromosome 8. Nature 439:331-335.

Palmour RM, Mulligan J, Howbert JJ, Ervin F (1997) Of monkeys and men: vervets and the genetics of human-like behaviors. Am J Hum Genet 61:481-488

Pevzner P, Tesler G (2003a) Genome rearrangements in mammalian evolution: lessons from human and mouse genomes. Genome Res 13:37-45.

Pevzner P, Tesler G (2003b) Human and mouse genomic sequences reveal extensive breakpoint reuse in mammalian evolution. Proc Natl Acad Sci U S A 100:7672-7677.

Regelson M, Eller CD, Horvath S, Marahrens Y (2006) A link between repetitive sequences and gene replication time. Cytogenet Genome Res 112:184-193.

Rogatcheva MM, Rund LA, Swanson KS, Marron BM, Beever JE, Counter CM, Schook LB (2004) Creating porcine biomedical models through recombineering. Comp Funct Genom 5:262-267.

Rogers J, Garcia R, Shelledy W, Kaplan J, Arya A, Johnson Z, Bergstrom M, Novakowski L, Nair P, Vinson A, Newman D, Heckman G, Cameron J (2006) An initial genetic linkage map of the rhesus macaque (Macaca mulatta) genome using human microsatellite loci. Genomics 87:30-38.

Rogers J, VandeBerg JL (1998) Gene Maps of Nonhuman Primates ILAR. Vol. 39 (2/3)

Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, Muzny D, Platzer M, et al. (2005) The DNA sequence of the human X chromosome. Nature 434:325-337.

Rudd MK, Willard HF (2004) Analysis of the centromeric regions of the human genome assembly. Trends Genet 20:529-533.

Rudd MK, Wray GA, Willard HF (2006) The evolutionary dynamics of alpha-satellite. Genome Res 16:88-96.

Samadashwily GM, Raca G, Mirkin SM (1997) Trinucleotide repeats affect DNA replication in vivo. Nat Genet 17:298-304.

Sanger F, Coulson AR, Barrell BG, Smith AJ, Roe BA (1980) Cloning in single-stranded bacteriophage as an aid to rapid DNA sequencing. J Mol Biol 143:161-178.

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB (1982) Nucleotide sequence of bacteriophage lambda DNA. J Mol Biol 162:729-773.

Schmutz J, Martin J, Terry A, Couronne O, Grimwood J, Lowry S, Gordon LA, et al. (2004) The DNA sequence and comparative analysis of human chromosome 5. Nature 431:268-274.

Schueler MG, Dunn JM, Bird CP, Ross MT, Viggiano L, Rocchi M, Willard HF, Green ED (2005) Progressive proximal expansion of the primate X chromosome centromere. Proc Natl Acad Sci U S A 102:10563-10568.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF (2001) Genomic and genetic definition of a functional human centromere. Science 294:109-115.

Schwimmer B (1998) www.umanitoba.ca/anthropology/courses/121/primatology/catarrhine.html Accessed June 9, 2006.

Shaffer LG, Lupski JR (2000) Molecular mechanisms for constitutional chromosomal rearrangements in humans. Annu Rev Genet 34:297-329.

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78-88.

Shizuya H, Birren B, Kim UJ, Mancino V, Slepak T, Tachiiri Y, Simon M (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc Natl Acad Sci U S A 89:8794-8797.

Slamovits CH, Rossi MS (2002) Satellite DNA: Agent of chromosomal evolution in mammals. A review. J Neotrop Mammal 9(2):297-308.

Spence JM, Critcher R, Ebersole TA, Valdivia MM, Earnshaw WC, Fukagawa T, Farr CJ (2002) Co-localization of centromere activity, proteins and topoisomerase II within a subdomain of the major human X alpha-satellite array. Embo J 21:5269-5280.

Stankiewicz P, Lupski JR (2002) Molecular-evolutionary mechanisms for genomic disorders. Curr Opin Genet Dev 12:312-319.

Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR (2004) Serial segmental duplications during primate evolution result in complex human genome architecture. Genome Res 14:2209-2220.

Stanyon R, Fantini C, Camperio-Ciani A, Chiarelli B, Ardito G (1988) Am J Primatol 16:3-17.

Thomas JW, Schueler MG, Summers TJ, Blakesley RW, McDowell JC, Thomas PJ, Idol JR, Maduro VV, Lee-Lin SQ, Touchman JW, Bouffard GG, Beckstrom-Sternberg SM, Green ED (2003) Pericentromeric duplications in the laboratory mouse. Genome Res 13:55-63.

Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, Haugen E, Hayden H, Albertson D, Pinkel D, Olson MV, Eichler EE (2005) Fine-scale structural variation of the human genome. Nat Genet 37:727-732.

Van der Kuyl AC, Dekker JT, Goudsmit J (1996) St. Kitts green monkeys originate from West Africa: genetic evidence from feces. Am J Primatol 40:361-364.

VandeBerg JL, Williams-Blangero S (1996) Strategies for using nonhuman primates in genetic research on multifactorial diseases. Lab Anim Sci 46:146-151.

Venter JC, Smith HO, Hood L (1996) A new strategy for genome sequencing. Nature 381:364-366.

Ventura M, Archidiacono N, Rocchi M (2001) Centromere emergence in evolution. Genome Res 11:595-599

Ventura M, Mudge JM, Palumbo V, Burn S, Blennow E, Pierluigi M, Giorda R, Zuffardi O, Archidiacono N, Jackson MS, Rocchi M (2003) Neocentromeres in 15q24-26 map to duplicons which flanked an ancestral centromere in 15q25. Genome Res 13:2059-2068,

Weinberg J, Stanyon R, Jauch A, Cremer T (1992) Homologies in human and Macaca fuscata chromosomes revealed by in situ suppression hybridization with human chromosome specific DNA libraries. Chromosoma 5-6:265-270.

Woo SS, Jiang J, Gill BS, Paterson AH, Wing RA (1994) Construction and characterization of a bacterial artificial chromosome library of Sorghum bicolor. Nucleic Acids Res 22:4922-4931.

Yunis JJ, Prakash O (1982) The origin of man: a chromosomal pictorial legacy. Science 215:1525-1530.

Yunis JJ, Sawyer JR, Dunham K (1980) The striking resemblance of high-resolution G-banded chromosomes of man and chimpanzee. Science 208:1145-1148.

Zhang L, Lu HH, Chung WY, Yang J, Li WH (2005) Patterns of segmental duplication in the human genome. Mol Biol Evol 22:135-141

Glossary

Below I have provided definitions for a selected number of genetic terms. The definitions were chosen (and directly quoted) from two different reference sources (Baxevanis and Ouellette 2005) and (http://www.genome.gov/glossary.cfm).

acrocentric chromosomes - Chromosomes with the centromere at or near one end of the chromosome.

BAC - A bacterial artificial chromosome is a DNA construct, based on a fertility plasmid, used for transforming and cloning in bacteria, usually E. coli. Its usual insert size is 150 kbp, with a range from 100 to 300 kbp.

chimeric read – In sequencing, a read containing sequence from two non-contiguous regions of the target or vector. Chimeric reads can be the result of multiple inserts ligating into the same vector during library construction, or sequence from a mixture of two clones that have regions in which each of the clones is more obvious.

chromosome - One of the threadlike "packages" of genes and other DNA in the nucleus of a cell. Different kinds of organisms have different numbers of chromosomes. Humans have 23 pairs of chromosomes, 46 in all: 44 autosomes and two sex chromosomes.

clone - A piece of DNA introduced inside a vector.

consensus – In alignments, the base most likely to occur at any given position.

deletion – A mutation in which one or more bases is lost from a given region of a chromosome.

duplication - A mutation which leads to the production of one or more copies of any piece of DNA, including a gene or even an entire chromosome.

electroporation – A method used to introduce plasmid into cells. Cells are poured in with a solution with plasmids. An electrical impulse is used to create transient pores in cellular membranes, thereby increasing the efficiency of uptake of plasmids from solution.

genetic map – A chromosome map of a species that shows the position of its known genes and/or markers relative to each other, rather than as specific physical points on each chromosome.

genome – All of the DNA found within each of the cells of an organism. Eukaryotic genomes can be subdivided into their nuclear genome (chromosomes found within nucleus) and their mitochondrial genome.

GSS – Genome survey sequence. This DDBJ/EMBL/GenBank division contains genomic sequences obtained from the following types of data: random 'single-pass-read' genome survey sequences, single-pass reads from cosmid/bacterial artificial chromosome/yeast artificial chromosome ends, exon-trapped genomic sequences, and Alu PCR sequences.

homologous – In phylogenetics, particular features in different individuals that are descended genetically from the same feature in a common ancestor.

identity – A quantitative measure of how related two sequences are to one another, assessed as the total number of exact matches in a pairwise sequence alignment.

insertion – A mutation in which one or more bases are inserted into a region of DNA.

karyotype - The chromosomal complement of an individual, including the number of chromosomes and any abnormalities. The term is also used to refer to a photograph of an individual's chromosomes.

library – In sequencing, a collection of insert-containing clones. Sequencing libraries are created from sequencing vector and a set of inserts obtained by fragmentation of a larger piece of DNA.

nucleotide – The basic component of both DNA and RNA. Nucleotides consist of base (adenine, cytosine, guanine, or thymine), a sugar molecule, and a phosphoric acid molecule.

orthologous – Homologous sequences are said to be orthologous when they are direct descendent of a sequence in the common ancestor, that is, without having undergone a gene duplication event.

paralogous – Homologous sequences in two organisms, A and B, that are descendants of two different copies of a sequence that has been created by a duplication event in the genome of the common ancestor.

physical map – A genome map showing the exact location of genes and markers. The highest-resolution physical map is the DNA sequence itself.

plasmid – A circular or linear, self-replicating piece of bacterial DNA.

repetitive DNA – DNA sequences of variable length that occur in multiple copies in the human and other eukaryotic genomes.

restriction fingerprint – The sizes of the DNA fragments resulting from an endonuclese digestion of the piece of DNA of interest.

phylogeny - The origin and evolution of a set of organisms, usually a set of species.

phylogenetics - Study of evolutionary relatedness among various species or other entities that are believed to have a common ancestor.

shotgun sequencing – A sequencing method in which the DNA to be sequenced is broken into many small fragments. The fragments in turn are sequenced individually; based on overlaps between the individual sequences, the pieces can be reassembled and the original sequences can be deduced.

similarity – A quantitative measure of how related two sequences are to one another, usually assessed as the total number of identities and conservative substitutuions in a pairwise sequence alignment. Similaritry does not imply homology

synteny – The preserved order of genes between related species. The order of genes is generally preserved best between tightly related species.

YAC – A yeast artificial chromosome is a vector used to clone large DNA fragments (larger than 100 kb and up to 3000 kb). It is an artificially constructed chromosome and contains the telomeric, centromeric, and replication origin sequences needed for replication in yeast cells.